DOCUMENT RESUME

ED 061 776 EM 009 937

AUTHOR	Lohnes, Paul R.
TITLE	Planning for Evaluation of the LRDC Instructional Model.
INSTITUTION	State Univ. of New York, Buffalo.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.; Pittsburgh Univ., Pa. Learning Research and Development Center.
REPORT NO	P-5-0253-1972-5
BUREAU NO	BR-5-0253
PUB DATE	72
NOTE	106p.
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	Academic Achievement; Conceptual Schemes; *Educational Theories; *Elementary Education; *Evaluation Criteria; Individual Differences; *Instructional Design; Intelligence Differences; *Summative Evaluation
IDENTIFIERS	IDE; Instructional Design and Evaluation; Project TALENT

ABSTRACT

The instructional model (IM) that is the basis of this evaluation resulted from the merger of two major Learning Research and Development Center (LRDC) projects--the Primary Education Project (PEP) and the Individually Prescribed Instruction (IPI) project. This paper examines the Center's publications relative to the problem of evaluation of such an IM and suggests new directions for summative evaluation. It emphasizes the importance of organizing ideas (educational theories) in evaluative research. Seven requirements of a theory of educational criteria are examined and a model is developed for combining information on student entering behaviors, educational treatment, and resulting student achievement. A list of 32 recommendations summarizes the goals for the future refinement of the IM. (JY)

OE-NCERD
EM
5-0253

LEARNING RESEARCH AND DEVELOPMENT CENTER

1972/5

PLANNING FOR EVALUATION OF THE LRDC INSTRUCTIONAL MODEL

PAUL R. LOHNES

ED 061776

EM 009 937

PLANNING FOR EVALUATION OF THE LRDC INSTRUCTIONAL MODEL

Paul R. Lohnes
State University of New York at Buffalo

1972

U.S. DEPARTMENT OF HEALTH, EDUCATION
& WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

The preparation of this essay was supported by the Learning Research and Development Center supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the Office of Education and no official endorsement should be inferred.

TABLE OF CONTENTS

Page

Foreword iii

Evaluation of Instructional Models 1

Delineating the LRDC Instructional Model 5

A Theory of Educational Criteria 15

Dimensions of Educational Treatments 35

Organismic Inputs 45

Tests for Curriculum Evaluation 60

Data Analysis 68

Data Bases 85

Summary of Recommendations 93

References 98

FOREWORD

I am pleased that the LRDC Publications Committee decided to add this monograph by Paul R. Lohnes to the Center Series, since I believe it represents a major contribution to the evaluation literature. Paul's assignment was to examine Center publications related to the problem of evaluation and to suggest new directions for summative evaluation. As he has indicated, "the tremendous accomplishments of the LRDC team in the area of formative evaluation have been underplayed in this manuscript only because the assignment was to shift attention to summative evaluation." This he has done in a remarkably insightful fashion.

One important contribution here is in reminding us that evaluative research, like any research, must have organizing ideas (theories) if it is to be maximally useful. Not only did he identify that need, but he contributed to meeting it by suggesting a theory of educational criteria, by pointing out the importance of assessing the degree of implementation of new educational programs, and by proposing a model for combining information on student entering behaviors (what he calls organismic inputs), educational treatments, and resulting student achievement.

One small caution to the reader is that Paul wrote this monograph at some distance from the Center. This was deliberate, but it means that his descriptions of LRDC and its instructional products are based upon perceptions gained from what he read. Any shortcomings then in his descriptions of "the LRDC model" are ours, or at least our literature's. What the reader will find in this monograph is an exciting trip through the evaluation problem and one man's perception of what LRDC is and what it is trying to do.

William W. Cooley
Co-Director, LRDC

PLANNING FOR EVALUATION
OF THE LRDC INSTRUCTIONAL MODEL

Paul R. Lohnes
State University of New York at Buffalo

Evaluation of Instructional Models

Intellectual preparation for the evaluation of instructional models would seem to involve at least the following requirements:

1. The need for delineation of an adequate theory of intelligence in particular and personality in general to organize the criterion domain. Long lists of specific behavioral objectives, although essential for the engineering of a curriculum, are not a substitute for the required theory. Indeed, such lists may impede the planning of a general evaluation, as a case of the adage about the trees making the forest difficult to see.

2. The need for a theory of instruction to organize the treatment domain. What are the significant dimensions on which instructional models differ? It should be possible to measure the degrees of realization of various stimulus-condition desiderata under different instructional models, rather than simply count them as present or absent. Treatments are more realistically viewed as differing in degrees of emphases, rather than in absolute kind.

3. The need for a theory of child development and achievement motivation to organize the domain of organismic inputs. Input individual differences necessarily account for the majority of criterion performance variance in schooling situations. This phenomenon is the principal characterizer of the problem of evaluating instructional models that makes it so much more difficult to accomplish that task than it is to do evaluations of many other systems and products. For this reason the medical

evaluation model is less than appropriate. LRDC's diagnostic-remediation approach to organismic variation in readiness for reading is exciting, but it doesn't alter the fact that while time is being expended in the correction of perceptual-motor deficits for some children, other children who don't have those deficits have the opportunity to forge ahead in learning to read.

4. The need for selection of measurement conventions for all three domains of variables. In some cases new test and inventory items will be required, although the invention of new item forms is extremely difficult. In most cases the appropriate scaling conventions could be provided by factors which are linear combinations of selected items.

5. The need for delineation of data analysis conventions appropriate to the problem of accounting for the criterion vector variance. The methodology will have to be multivariate, correlational, and heuristic. The ways in which covariance analysis can be obfuscating and misleading must be understood.

6. The need for research designs that establish credible data bases. Because treatment variables are almost inevitably correlated with organismic inputs in schooling situations, obtaining convincing contrasts is more difficult than in other researches. Classical experiments, with random assignments of subjects to treatments, are just about out of the question. Another issue to be resolved is the unit of analysis in evaluation studies. Should it be the school class or the individual pupil?

7. The need for a theory of educational development that takes a long view into the varied futures of children and illuminates the possible value judgments regarding what is good for them, without falling into the trap of a narrow dogmatism that is unsuitable for a free society and that will alienate more clients than it can possibly please.

In short, in its evaluation presentations LRDC has to encourage good thinking on the part of its clients regarding what an instructional model can properly be asked to accomplish, as opposed to what only God or evolution could have done differently.

The LRDC instructional model relies heavily upon research-established learning hierarchies. It will be emphasized that empirical hierarchies are demonstrations of possible ladders to generalized abilities, but are not necessary ladders. Since flexibility is the outstanding characteristic of human intelligence, there should be alternative progressions to particular objectives. Showing that a particular progression works is not itself sufficient to proclaim a summative evaluation accomplished, although it is impressive and useful information for other purposes.

It will be argued that the shape of a criterion vector distribution, including its orientation, is more meaningful than its location, even when treatment variation is associated with location variation. Anova enthusiasts give too little thought to the implications of the usual enormous overlaps of criteria distributions that are located "significantly different." Also, it can be argued that the social utilities of effects on outliers, both very low achievers and very high achievers, are much greater than of effects on central tendency. Finally, if it is established that instructional models differ in degrees rather than in kind, then a locations logic is invalidated anyway.

Theory simplifies reality in order that human thought may grasp the fundaments of experience. Theories deal in generalizations relating concepts from different domains of observation. A theory of the evaluation of instructional models should relate concepts from the domains of organismic inputs, treatments, and achievements. It should yield probabilistic propositions that allow prediction and control of future experiences of children in school classes. Arguments about causality are bootless in the evaluation context because of the spiralling of influence among the domains. The measurements required by the theory should be practical, which is to say parsimonious and economical. The data bases and analyses selected should facilitate sound and sharp decision-making by those responsible for the educational policies of the communities, states, and nation.

Delineating the LRDC Instructional Model

At the Learning Research and Development Center variety in activities is as evident as singleness of purpose. Currently there are 24 projects underway. During 1970 several projects were phased out and several new ones were started. Animals and computers are the subjects of some of the projects. Those projects for which the subjects are school children and teachers vary considerably in their degrees of direct and immediate relevance to educational practices in American schools. Although there is a common approach to educational innovation that now organizes some of the projects, an approach that can be delineated as the LRDC Instructional Model (IM), an evaluation of this IM must be less than a total evaluation of LRDC. The current version of the IM could be evaluated as no great improvement over existing competing instructional models and it could still be that LRDC is the R&D shop most likely to emerge with a superior IM in time. One or more of the projects that now seem remote from the thrust of implementation of the current IM might contain the seed ideas for a breakthrough.

Still, LRDC is intended to have a practical, engineering emphasis to its work, and after seven years of existence it is reasonable to inquire what of practical value has been achieved. Attention will be directed to the evaluation of an entire system of elementary education because that is what LRDC appears to be striving toward. They have designed most of a system and are close to the completion of a prototype "new school." However, evaluation of the total system could be disappointing and it still could be that some modules of the system are outstandingly successful new designs. It seems reasonable to expect that some components of a new system would perform better than others, and that even within modules there might be specific units of great value and some of near worthlessness under unanticipated field conditions. Any evaluation of a new system is likely to yield detailed feedback information to its engineers. Attention here will be directed to the information about the overall performance of the LRDC system qua system that an evaluation should provide to prospective clients. The concern will be with evaluation of the LRDC "new school" in behalf of clients who are seeking a better elementary education system for the children they have in trust.

The LRDC IM is the offspring of the marriage of the Primary Education Project (PEP) and the Individually Prescribed Instruction project (IPI). This recent merger of the two "new school" thrusts in a single new program has inspired recognition that LRDC now has a coherent instructional model that provides at least the design sketch and some, if not all, of the blueprints of a new elementary education system. The IM still lacks a handy mnemonic and it has yet to be delineated in detail in any one document. Nevertheless it is tangible and unmistakable in the increased cooperation and integration among several projects and is documented, albeit diffusely, in the 1970 Annual Report and the 1970 Program Plan and Budget Request. Several recent staff publications (e.g., Glaser & Nitko, 1971) take the existence of the IM for granted. For present purposes the following brief review of the principles and components of the LRDC IM will have to suffice to characterize what is to be evaluated.

The first principle on which the IM is built is that the curriculum is to be adapted to the individual pupil in terms of entry point, pace, density of drill, and sequence of units. Through intensive monitoring of what the pupil can do, using a variety of placement, diagnostic, and criterion tests, the curriculum operators should know precisely what he can be required to attempt next. The principle which makes it possible to prescribe next learning objectives from test information of present capabilities is that there are hierarchies of abilities which map out the subcompetencies at a next lower level on which competencies at a level can be constructed. Each of these hierarchies is a pyramid that has a base of a large number of simple learnings and an apex of a single generalized ability. There are many "pyramids" within each completed hierarchy, in that any subordinate learning is also superordinate to a set of subordinate learnings on which it is built (except, of course, those input "aptitudes" that provide the base level of the hierarchy). Prescription is possible because knowing where a pupil is on this map indicates his opportunities for a lateral or upwards move. Operators may have to exercise judgment when two or more moves are available, however. There is a strong emphasis in the IM on preparing the pupil to make these judgments himself, becoming his own "operator."

Human intelligence is omniverous, if never omniscient, and the variety of generalized abilities for which learning hierarchies could be constructed is virtually endless. The IM adopts a Draconian solution in the principle of preselected behavioral objectives which provide the polar constellation for a curriculum. These curriculum objectives are the constant goals for all pupils, although they must inevitably remain always over the horizon for some. They are adult wisdom imposed on neophytes. So it must be; so be it. This is an IM for elementary education.

Since Thorndike formulated the Law of Effect, all instructional models have given consideration to the provision of reinforcements for learning, but the LRDC IM is noteworthy for its adherence to a principle of very active, responsive participation of the pupil in his schooling and of a high density of feedback and reinforcement. Quizzes embedded within units contribute to this reinforcement, and in some curricula there is heavy emphasis on contingency management by traveling teachers. The IM looks to pupil knowledge of objectives and the experience of success as sources of pupil motivation.

These, then, would seem to be the principles on which the IM is founded:

1. Individualization.
2. Learning hierarchies.
3. Behavioral objectives.
4. Reinforcement and motivation.

If there is one technique that dominates the IM it is testing: testing for placement, testing for diagnosis, testing for prescription, testing for reinforcement, and testing for achievement. There is so much testing in LRDC curricula that it is the most likely place for problems to arise. It has been pointed out that both pupils and teachers can develop work habits that abuse the tests in ways which jeopardize the entire instructional system (Reynolds, Light, & Mueller, 1970). A thesis of the analysis of this essay will be that the LRDC testing is incomplete in ways that may contribute to partial understanding of the pupil and his schooling. The analysis will also suggest that learning hierarchies are perhaps misinterpreted in some of the LRDC literature, and will pose some questions about preselected behavioral objectives. On the whole, though, the IM has the advantage of a foundation in a selection of psychological principles capable of giving clear direction to curriculum construction.

Curricula have been constructed and tested for most of the crucial learning areas and levels of elementary education. In fact, some of the most promising work is focused on what used to be considered the preschool ages of three to five. LRDC's "new school" will stretch the elementary education process to incorporate children two or three years younger than is customary. In these early years they hope to engender that readiness for formal language and mathematics instruction that is so sadly lacking in many first-graders. A remark by Glaser in a staff seminar to the effect that PEP is "creating intelligence" describes what is happening. Those basic skills and concepts which are measured in intelligence tests for preschoolers and in reading readiness batteries are being taught to children who lack them. These include perceptual-motor skills, language concepts, logical processes such as classification, and concepts of number. Placement and diagnostic tests keyed to the specific behavioral objectives have been developed for all units of instruction. Besides the curriculum units, the PEP projects also develop school and classroom organizations, teacher and staff training procedures, and approaches to relations with parents.

The Individualized Perceptual Skills Curriculum (Rosner, 1969) illustrates the careful intermix of basic research and creative design of instruction that characterizes PEP projects. Starting with the psychological principle that readiness for learning to manipulate abstractions in language and math is based at least in part on facility in processing concrete sensory information, Rosner refined an elaborate testing procedure for diagnosing perceptual-motor dysfunctions. Armed with evidence of widespread occurrence of specific deficits, he undertook an analysis of the precise decoding, encoding, and intersensory integrative skills assumed by primary reading and other first-grade instructional programs. There followed the highly creative activity of inventing exercises to teach these precise skills, and the tedious trials with pupils required to discover what would work and how to improve the workable. In Table 1, the list of the units developed for visual-motor training shows the detailed nature of the analysis and the engineering. Equally well-developed unit-sequences cover the auditory-motor, general-motor, and integrative areas.

This is the kind of frontal assult on the need "to teach the skills and concepts that underlie intelligent behavior" (Resnick, 1967) that PEP has substituted for the educator's usual posture of deploring the low

Table 1. Visual-motor program, 1969-70

Topics	Number of Objectives
Unit 1--Superimposition	11
Unit 2--Direct match from model	12
Unit 3--Match from abstract model to concrete representation	17
Unit 4--Match from concrete model to abstract representation	16
Unit 5--Match from concrete model to partial abstract representation	6
Unit 6--Combination of visual-motor processes	8
Unit 7--Letter identification	16
Total number of objectives	86

intelligence of a child fated to be a failure in school. Educators who want to evaluate the LRDC IM really should start by surveying the detailed knowledge of learning and instruction that has accrued to LRDC by hard work under the direction of the instructional model.

The PEP classification program provides another example of the sort of highly articulated curriculum components LRDC has achieved. There are three levels in the program, with several units for each level, and several to many behavioral objectives for each unit. The lowest level contains two units on matching skills and five units on discrimination skills. The middle level has two units on prepositional statements and two units on identity statements. The highest level consists of three units, on advanced color, size, and shape discrimination, on functional categories, and on category naming. All told, there are 84 specific behavioral objectives for which precise definitions, instructional procedures, and tests exist. PEP's reading and mathematics programs are too detailed for description here, although there will be some analysis of the mathematics hierarchies later. Noteworthy progress has been made in bridging from the PEP to the IPI program sequences in these areas, although the exact relationships between correlated units of the programs are often hard to discern from the literature.

The oldest, most extensive, and most widely deployed of LRDC's curricula is the IPI mathematics program for grades one through six. Somewhat less firmly established is the IPI reading program for these school years, but both sequences are engineered on the principles of the IM. Versions of IPI math are in use in hundreds of schools. A less-well-known and more easily discussed curriculum is the IPI Individualized Science. The aim of this curriculum project has been stated as "no less than the development of a complete individualized science learning system." Five macro-goals have been set for the system:

(1) self-direction, (2) co-evaluation, (3) positive affect for science, (4) inquiry abilities, and (5) scientific literacy. The classic rhetoric of general education has been invoked to justify this program: "Our overriding aim is to offer a program of elementary-school science education that is relevant, meaningful, and of greatest benefit to the individual. We shall know that this aim has been achieved when we see that the student in Individualized Science develops positive feelings about science and about himself through his active involvement in self-directed, self-fulfilling, satisfying science learning" (Klopfer, 1971, p. 11). Such rhetoric poses enormous measurement problems for the evaluator if it is taken seriously, but it is noble coinage, nevertheless.

The science learning system has been subdivided into levels corresponding to school years, and into units, usually three per level. All five macro-goals are considered to apply to all levels and even to all units. The units themselves are far less prescriptive than those of IPI mathematics, and might better be considered resource units than instructional units. "To make possible the realization of all five goals, our principal strategy is to make available a large variety of learning resources for the student's use in the program" (Klopfer, 1971, p. 8). There is a contract with Imperial International Learning (IIL) to package completed units for production and distribution. It is the finished units from IIL that the Research for Better Schools (RBS) regional lab gets for release to its cooperating schools, for example. Thus the completed science units might be viewed as a "product" of LRDC that is open to evaluation.

It is too simplistic to view the units separately, however. Evaluation of separate units, even in the field, must be construed as engineering, or formative, evaluation, rather than summative evaluation intended to establish the competitive advantages of curriculum programs. In a memo, Champagne and Klopfer (10/5/71) give an interesting, detailed flow-chart of the formative evaluation steps a unit undergoes before it becomes a finished product. There is review of each "resource" created for inclusion in a unit by the science staff, including tryout of manipulative aspects with children. When all resources are ready, there is review of the integrated unit by the measurement and evaluation staff prior to prototype production. Then there is a tryout of the prototype at the Oakleaf School, leading to a review before production by IIL of the field testing version. Field testing is done in RBS schools. There is review of the field test results involving LRDC, RBS, and IIL staff members. Thus, there are at least four occasions for review and revision, involving expanding datasets and increasing collaborations, prior to attainment of a "product." All of this contributes to knowledge of how each resource works in conjunction with the rest of the unit to nurture development on each of the five objectives of the program. Resources can't be evaluated properly outside the framework of this working together, so the unit becomes the smallest feasible micro-treatment available for evaluation. However, it will be argued later in this essay that all we know regarding the development in human personality of general traits, sets, styles, and habits coheres to suggest that reliable changes in such personality dispositions will be associated only with prolonged and cumulative treatments. That is, education is not usually traumatic. Psychological theory prompts the assertion that the evaluation of units is likely to be formative in purpose, and that levels of curricula representing about a year's work will be the smallest feasible treatments for summative evaluation. It might be argued that the entire science learning system in several levels should be evaluated as a whole to meet the information needs of prospective adopters.

The primary education process developed in O. K. Moore's Responsive Environments Laboratory, as well as the "Stepping Stones to Reading" program invented by P. M. Kjeldergaard, and the computer-based spelling and math testing programs are examples of curriculum development projects at LRDC that are considered in this essay to be independent of, and not derivative from, the IM under discussion. Just within the PEP and IPI areas the IM has generated a variety of curriculum components that is difficult to cognize. There is need for LRDC to produce a master chart of what has been developed, organized into curriculum components within curricula and curriculum units subordinate to components. The developmental level for each unit should be shown, as well as the hierarchical relations among units. The behavioral objectives and internal learning hierarchy for each unit should be readily retrieved. All this would provide a master plan of the program for the "new school," to which evaluative information could be keyed. There would be flux in this "periodic table," and both the complexity and flux would make it clear that there can never be a simplistic or final evaluation of the LRDC Instructional Model and its derivative educational system.

A Theory of Educational Criteria

LRDC's strongest principle of operation in curriculum development may also be its greatest obstacle to a workable approach to summative evaluation. The principle of behavioral objectives, although absolutely necessary for the engineering accomplishments of the LRDC team, may have become an obsession capable of blinding true believers to the equally necessary principle of evaluation, which requires that school learnings be organized into generalized achievement traits. Behavioral objectives are multitudinous discriminations needed in the making of a curriculum. Achievement traits are parsimonious integrations needed in telling about a curriculum. The analytical mode for the creative act, but the synthetic mode for the persuasive dialogue about the created program.

People simply will not sit still long enough to be told the list of behavioral objectives of a curriculum and how well each is achieved, retained, and applied by pupils.

If learning hierarchies are to be a really convenient fiction they must be interpreted as blueprints for the building of generalized traits. Otherwise they may be temporarily helpful in the parochial environs of the R&D shop but dismally obstructive to the dialogue in the educational marketplace. The guide who persists in running on about where each block is in the edifice will soon lose rapport with the visitor who wants to know what the pyramid is for. There are always many ways to know something. The designer's way of knowing his thing does not serve the consumer's need to know. Where the designer sees process the consumer sees product. They see in different sets of questions and should be answered in different sets of concepts and generalizations.

Teaching and learning are interrelated in a process that requires careful, precise description, as discussed in the next section of this essay.

Education as the product of the teaching-learning process is an abstraction of the highest order of generalization. No one can say precisely what it is but masses of people desire it, although usually more for their children than for themselves. The most naive view is that education is defined by possession of diplomas, certificates, and degrees. These hallmarks of education have an objective reality that appeals to the researcher seeking a reliable criterion variable. Later, in a discussion of longitudinal data bases, it will be argued that evaluation should be planned to relate treatments to this and similar classes of objective career achievement variables. Most people would agree, however, to a sense of education in which it means the qualities of an educated person that qualify him for special prizes and places in real life. What are the qualities which comprise education? They have been variously described as "knowledges, information, skills, attitudes, appreciations, personal-social adaptability, interests, and work habits" (Tyler, 1951, p. 48) but primarily they are classes of knowings: knowing that's and knowing how's. Thus, the knowing of mathematics is a knowing that mathematics consists of a certain corpus of material and a knowing how to do mathematics acts.

The dictation of reading and mathematics goals to children is a prevailing given of elementary education, not an invention of LRDC. The LRDC instructional model is more concerned with showing the schools how to do their conventional work than it is with inquiry into what the goals of schools should be. It is perhaps for this reason that LRDC has been relatively inarticulate regarding the generalized traits its curricula are supposed to sponsor. Expecting LRDC's scientists to ask whether children need mathematics would be like expecting Boeing's engineers to ask whether the nation needs an SST.

The generalized traits that comprise the qualities of education are convenient fictions. They are figments of the imagination of scientists who are trying to understand the correlations between data on behaviors of children. To give them a fancier name, they are intervening variables that bridge between classes of data known to be related through correlations. These variables are the concepts of a theory of educational criteria. If curriculum scientists want to communicate effectively with the public they will have to establish in the public philosophy their language of concept names for educational criteria. This task will be easier if the scientists respect what the public already knows about the qualities of education and set about the refinement rather than the disruption of public knowledge. In this sense, respect for the ordinary language of education is essential.

Psychology, through its manifold avenues to the public's awareness and through the vehicle of educational psychology, has popularized several trait concepts to an extent that they cannot be ignored. Intelligence is the most noteworthy of these (and some of LRDC's pronouncements on it are perilously close to the "God is dead!" rhetoric), but spatial relations, mechanical reasoning, clerical speed and accuracy, and memory are well-known aptitudes. The language of vocational interest traits and of academic motivation is widely used. Educational psychology has tolerated the lexicon of separate fields of academic knowledges--reading, writing, arithmetic, and the like--for so long, despite the overwhelming correlation and factor analysis evidence against it, that the best to be hoped for is a refinement of public understanding and usage. In this regard, the test publishing agencies have promulgated some of the most obfuscated trait language in their grade-level achievement series. Since LRDC cannot wage internecine war against the practices of such as Educational Testing Service (ETS) and Science Research Associates (SRA), the approach has to be one of reconstruction of generalizations within the rubrics of established language conventions. (Fortunately, ETS can be counted on for helpful essays from the left hand in the forthcoming reports of research by K. Joreskog and G. W. Berglund into the SCAT-STEP factor structure, of which ETS says:

> The purpose of this study is to investigate stability and change over time of the factor structure of two ETS tests designed for use from grades 5 through 11, the School and College Ability Tests and the Sequential Tests of Educational Progress. Data from the total Growth Study sample are being analyzed. The approach involves fitting of alternative factorial models to test scores obtained at grades 5, 7, 9, and 11. Being taken into consideration are general factors across all tests, factors specific to individual tests across grades, and factors specific to particular grades. When completed, the research should be of interest to test developers and to primary and secondary school administrators and teachers.

Oh, yes, and to curriculum developers, too!)

The obvious reason for LRDC's lack of enthusiasm for the concept of intelligence is that so much of the voluminous research literature on the subject is concerned with intelligence in the role of predictor of mental development. In one place Glaser says: "Little use is made, at present, of measures of general intelligence or aptitude which have seemed difficult to relate to instructional decisions in the elementary school" (Glaser, 1968, p. 11). In another place he argues: "Global notions of general intelligence are obviously no longer useful scientific concepts for describing learner characteristics because such global measures tend to neglect and obscure specific individual differences. Rather, what is more important for instruction is to determine initial patterns of ability and competence that interact with learning. In the experimental and theoretical study of learning, resistance to discovering what may be hidden in error variance needs to be overcome" (Glaser & Nitko, 1971, p. 629). This LRDC view is correct when the purpose of measurement is curriculum placement. W. W. Cook said it years ago: "The best measure of what an individual should achieve in a given area is past achievement in that area" (Cook, 1951, p. 35). The essential role of intelligence measurement for statistical control in evaluation studies will be stated later. The present point is that in seeing so clearly that intelligence measurement is not required for prescription purposes, the LRDC team has in the past overlooked the need for intelligence measurement for criterion purposes. Glaser and Nitko are within reach of a great truth when they plead for "discovering what may be hidden in error variance." What is hidden in error variance (the residual variance available to curriculum researches after covariance adjustments for input differences in intelligence and aptitudes) is more intelligence: Intelligence is not only the best predictor; it is also the best criterion.

Bloom was wise to speak of "stability and change" in human intelligence. Intelligence is developed in people over the span of many years and through the interaction of constitutional and environmental circumstances. This development follows a negatively accelerated parabolic curve usually, so that growth spurts tend to occur early. Intelligence is the most noteworthy mental trait a child is growing in any school year, as well as the most noteworthy mental trait he possesses as he enters each school year. The marvel is that entry status and growth of intelligence are imperfectly correlated. Researchers have the opportunity to seek out the treatment correlates of unpredicted or residual development of intelligence. What is the evidence for intelligence as the best criterion of educational development? M. F. Shaycoft (1967) conducted a monumental retest study using Project TALENT subjects and tests. She had 7600 of the students TALENT tested in 1960 as ninth-graders retested when they were in twelfth grade. In their own analysis of Shaycoft's 96th-order correlation matrix (48 tests given twice), Cooley and Lohnes (1968, pp. 1-12 - 1-21) transformed to the rubrics of six factors of the TALENT battery, as proposed by Lohnes (1966). These consisted of one general factor and five residual, uncorrelated group factors (mathematics, English, hunting and fishing, visual reasoning, and perceptual speed and accuracy). The g factor was the most stable of the six over the high school years. Ninth-grade g correlated .77 with twelfth-grade g. (The next most stable relation was .65 between ninth visual reasoning and twelfth.) General intelligence as the most predictable criterion is a well-known phenomenon. But, the coefficient of determination for the g factor is only .59, indicating that about 40 per cent of the observed variance in twelfth-grade g is available as unexplained criterion variance. Since g is measured from the 48 tests with very high reliability, much of this residual variance is real and ought to be relatable to treatment variance.

More relevant to the LRDC elementary education program is a demonstration based on data of the USOE Cooperative Reading Studies Second-grade Phase. Lohnes and Gray (1972, in press) computed correlations among eight readiness tests and eleven first- and second-grade achievement tests for 3956 pupils from ten projects around the country. Sex and project effects were extracted as constants in a linear model, leaving the intercorrelations among residuals for further analysis. Only two principal components extracted more than unit variance. The first factor accounted for 53 per cent of the generalized variance of the 21 test residual scores. This factor was clearly an operationalization of general intelligence. The second factor accounted for eight per cent of the generalized variance and was interpreted as an auditory versus visual discrimination factor. A canonical correlation analysis of readiness tests versus achievement tests also resulted in a two-factor solution, with a g factor of readiness correlated .8 with a g factor of achievement. The second factors correlated only .3, and practically all of the 42 per cent redundancy of the achievement battery, given the readiness battery, was interpretable as persistence of general intelligence. It was argued that the USOE Reading Studies measurements were overwhelmingly saturated with general intelligence.

Table 2 gives details of the Lohnes and Gray canonical analysis which showed a g factor of the end-of-year achievement tests to be the most predictable criterion from a g-type predictor function of the eight beginning-of-first-grade readiness tests. This is an old story, albeit one not often enough told. The really interesting phenomenon is the correlations of achievement residuals after multiple partialling (Cooley & Lohnes, 1971, Ch. 7) of the eight readiness tests. Table 3 shows these for ten language arts tests (arithmetic computation was dropped from the analysis). All correlations between residual achievement scores were positive, making a dominant g-type factor inevitable (as much as factors ever are). Only one eigenvalue of the matrix was larger than unity, and the g factor associated with it accounted for 51 per cent of the generalized variance for residuals of the ten tests. Factor loadings (next-to-last row of Table 3) showed that the criterion g differed from the corresponding canonical analysis g (of Table 2) most in the lower loading of the Vocabulary test. This particular test was orally administered and was more like the readiness tests than the other achievement tests. The last row of Table 3 gives the squared multiple correlations for each achievement residual regressed on its nine fellows. If these are taken, as they often are, as approximate magnitudes for desirable communalities, it is apparent that a one-factor theory for the common variance in the residuals is quite adequate.

The intrusion of this detail in this essay is perhaps justified by the novelty and importance of the point under demonstration. Canonical correlation and multiple partial correlation were juxtaposed to show that when end-of-year achievement test scores were decomposed into independent,

Table 2. Canonical correlation factors (Lohnes & Gray, 1972, in press)

Readiness Tests	Factor Structure	
	g_R	s_R
1 Pintner-Cunningham	.84	-.19
2 Murphy-Durrell Phonemes	.81	.02
3 Murphy-Durrell Letter Names	.79	.48
4 Murphy-Durrell Learning Rate	.57	.34
5 Thurstone Pattern Copy	.60	-.06
6 Thurstone Identical Forms	.44	-.06
7 Metropolitan Word Meaning	.67	-.44
8 Metropolitan Listening	.54	-.39
% Variance Extracted	45	9
Achievement Tests (Stanford)	g_A	s_A
9 Word Reading, grade 1	.84	.36
10 Word Meaning, grade 2	.80	.11
11 Paragraph Meaning, grade 1	.84	.38
12 Paragraph Meaning, grade 2	.85	.12
13 Vocabulary, grade 1	.87	-.31
14 Spelling, grade 1	.71	.50
15 Spelling, grade 2	.69	.47
16 Word Study Skills, grade 1	.84	.20
17 Word Study Skills, grade 2	.79	.09
18 Language, grade 2	.76	.08
19 Arithmetic Computation, grade 2	.64	.11
% Variance Extracted	62	8
Redundancy to Readiness	.41	.01
Canonical Correlation	.81	.31

Table 3. Intercorrelations among residuals from multiple partialling of eight readiness tests (N = 3956)[a]

Test (Stanford)	1	2	3	4	5	6	7	8	9	10
1 Word Reading, grade 1		50	66	49	27	50	55	60	41	39
2 Word Meaning, grade 2			52	70	33	39	62	44	51	52
3 Paragraph Meaning, grade 1				54	31	50	52	55	42	43
4 Paragraph Meaning, grade 2					32	41	62	45	51	55
5 Vocabulary, grade 1						21	20	30	20	23
6 Spelling, grade 1							46	52	33	31
7 Spelling, grade 2								49	56	48
8 Word Study Skills, grade 1									44	37
9 Word Study Skills, grade 2										46
10 Language, grade 2										
Factor Structure	76	79	78	80	43	65	79	73	68	67
Multiple R^2	55	58	54	60	16	37	56	48	40	38

[a] A decimal point is implied immediately before each first digit in the table.

additive parts representing the predictable part from regression on a readiness battery and the error or residual part, the correlation matrices for the two parts were both positive manifolds, and each was dominated by a g-type factor. In practical terms these findings suggest that in a system of behaviors as highly interdependent and g-saturated as end-of-year achievement test performances, both the predictable (from input measures) and the unpredictable parts of those performances will be substantially interdependent and g-saturated. The close-to-the-chest technical implication is that a general, positive linear function of the residuals from multiple regression is probably going to be the best criterion factor for evaluation research.

General intelligence is by definition that latent factor of human personality that causes the degrees of positive correlation that always occur between all pairs of aptitude and ability measures. Spearman spoke of the "indifference of the indicator" to emphasize that every ability test has some g saturation and is to some extent a useful indicator of general intelligence. What researchers have to realize is that no one ability test adequately operationalizes general intelligence. A battery of ability tests is required, and the concept has to be operationalized by an appropriate multivariate statistical analysis that constructs a linear function as a g score. Whether enthusiasts for the concept of general intelligence should be permitted to describe such a g-type factor by their beloved concept name may be debatable, but the position taken here is that communication in the educational marketplace can be facilitated by such usage. "Intelligence" rolls off the tongue and sounds in the ear nicely, and there is a grandeur to the notion of "freeing intelligence through teaching" (Murphy, 1961). At least those among the LRDC team who find it acceptable to speak of striving "towards humanistic goals through behavioral objectives" (Beck, 1970) ought to be responsive to this conceptualization.

Some LRDC literature implies that only published standardized tests are haunted by the succubus "intelligence," while parochial criterion-referenced tests escape infestation. The empirical evidence on intercorrelations of LRDC curriculum tests and their cross-correlations with published tests is scanty and inconclusive, although what is available does suggest that modest-to-moderate positive coefficients prevail. Some of the evidence will be reviewed later in the discussion of data bases. LRDC should conduct factor analytic studies of its curriculum tests to inform the public regarding how they relate to popular trait concepts. Scalogram and simplex analyses (e.g., Boozer, 1970) are significant formative evaluation tools but they require the supplementary support of conventional correlation analyses when the criterion-referenced tests are pushed forward for summative evaluation purposes. LRDC acknowledges that in its tests "the groups of sequenced objectives are assumed to require varying degrees of mastery of the same intellectual skill" (Boozer, 1970, p. 114). Identification of the general intellectual skills is not too much to ask. If it is established that LRDC's curriculum tests have unusually high amounts of unique variance, this may actually reduce their relevance for summative evaluation.

PEP is one program at LRDC that has unabashedly acknowledged intelligence as a curriculum criterion. The PEP program has been designed to teach the skills that are tested in individual intelligence tests for children. The PEP team is prepared to use IQ gains as criterion performance evidence. In their 1968-69 report, they showed that 59 randomly selected kindergarten children who were tested twice (October 1968 and May 1969) with the Stanford-Binet had a mean IQ gain of 5.3 points, which they said "indicates that the PEP program has a significant impact on the children's general intellectual performance" (Wang, Resnick, & Schuetz, 1970, p. 7). To the PEP team the present argument must appear supercilious.

What LRDC has replaced generalized traits with in its theorizing is learning hierarchies. Hierarchies, rather than traits, are chosen to explain the structure of cognitive behaviors. This approach apparently follows R. Gagné's theory of a hierarchy of learning sets mediating a complex criterion performance (Gagné & Paradise, 1961). Individual differences in amounts and kinds of available relevant knowledge are said to be the source of differences in rate of achievement. These underlying knowledges are of both the knowing that and the knowing how types, and they are organized in a pyramidal structure, in which subordinate sets mediate transfer to higher level sets. The lowest level learning sets are much like what psychometricians call aptitudes. Thus, acquisition of required specific knowledges is dependent on the mediation of appropriate aptitudes, and in turn the acquisition of a complex ability or higher mental process depends primarily on transfer from specific knowledges. How many levels exist in a pyramid and what the building blocks are at each level are matters for research to establish. At LRDC this theory has been worked for curriculum prescription and measurement of student progress:

> Well-defined sequences of progressive, behaviorally defined objectives in various subject areas need to be established as guidelines for setting up a student's program of study. The student's achievement is defined by his position along this progression of advancement (Glaser, 1968, p. 5).

Anyone who reads broadly in the LRDC publications must be impressed by the sheer volume of detailed specification of empirically validated learning hierarchies that has been achieved. More will be said in praise of this scientific knowledge of human learning processes in the discussion of treatment variables.

An example of a learning hierarchy analysis is provided by the PEP Beginning Mathematics Program. There are 104 behavioral objectives progressively sequenced in 14 units. Figure 1 shows the complicated, nonlinear relations among the units, and presumably indicates that there are alternative paths by which the student can travel to unit 14 (and some other termini). Figure 2 is a mapping of the relations among the nine behavioral objectives of the first unit. Several issues need to be considered regarding this and similar learning hierarchies produced by LRDC:

1. What is the student achieving as he progresses along one of these curriculum sequences? What is the name of his acquisition? Presumably he is developing a complex ability or higher mental process that can be delineated as a trait and designated by an appropriate trait name.

2. Why isn't the hierarchy map more pyramidal? Why hasn't it a single end-point on which all paths converge? If more than a congeries of associations and habits is under development within a unit or an entire curriculum component, the map ought to show levels of fewer and more comprehensive, generalized knowledges as it is read upward. Of course, if there is no summit there is no mountain or pyramid to be named by a trait concept.

3. Doesn't the existence of alternative paths make the student's position on the map an ambiguous measurement of his achievement? Also, as will be described later, serious decay of learning occurs in an LRDC curriculum, as in any curriculum. If a student has forgotten what he learned previously, may not his present position merely show where he is "lost" in a maze?

4. Isn't it possible that other sequences of learnings that lead to the same summit ability could be created for children, or even invented by them? The PEP mathematics program has considerable overlap with the IPI mathematics curriculum, for example. It is well known that there are many ways to teach beginning reading, and LRDC has developed or

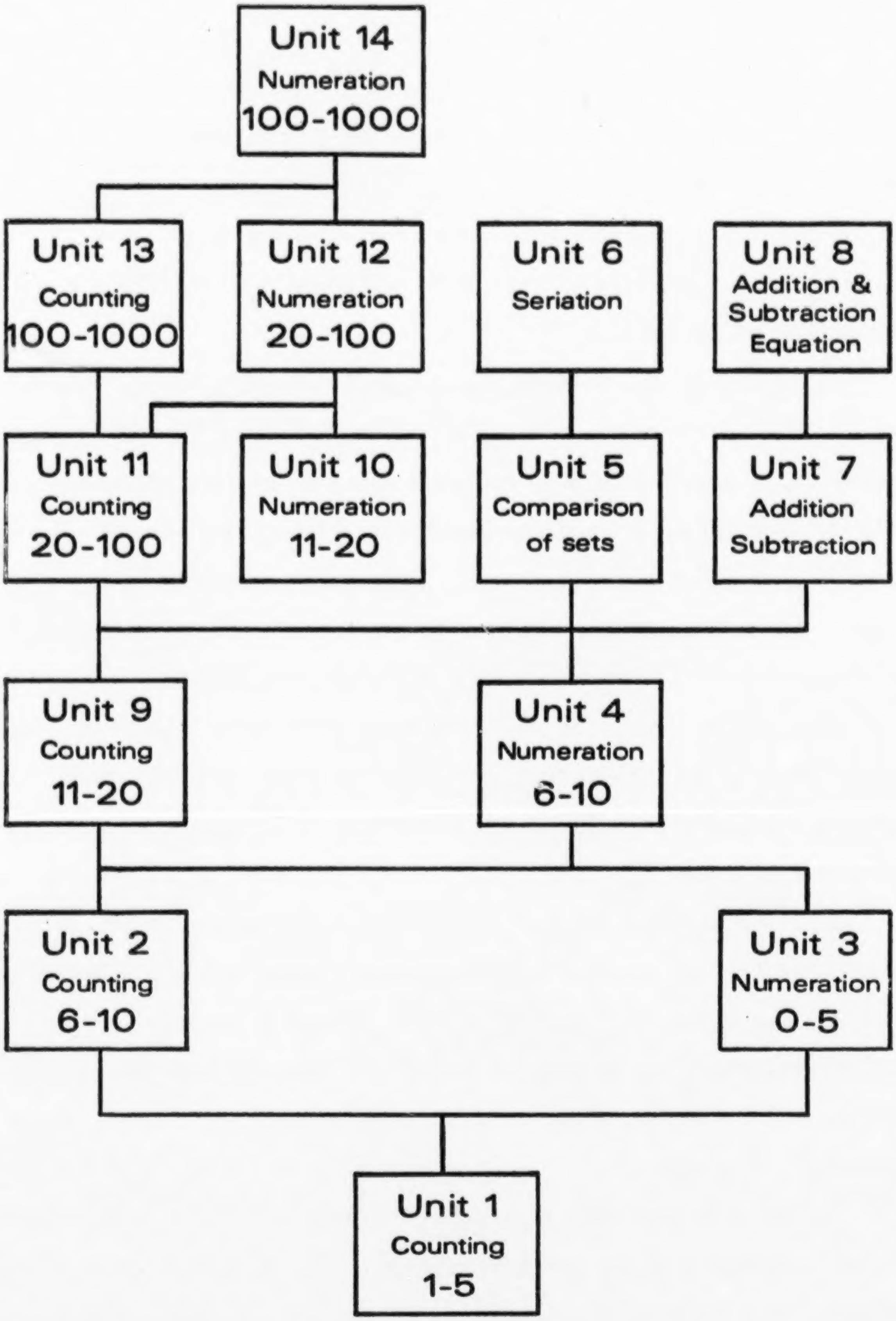

Unit 14
Numeration
100-1000
Unit 13
Counting
100-1000
Unit 12
Numeration
20-100
Unit 6
Seriation
Unit 8
Addition &
Subtraction
Equation
Unit 11
Counting
20-100
Unit 10
Numeration
11-20
Unit 5
Comparison
of sets
Unit 7
Addition
Subtraction
Unit 9
Counting
11-20
Unit 4
Numeration
6-10
Unit 2
Counting
6-10
Unit 3
Numeration
0-5
Unit 1
Counting
1-5

Figure 1
Curriculum Sequence Information for Quantification Units 1-14

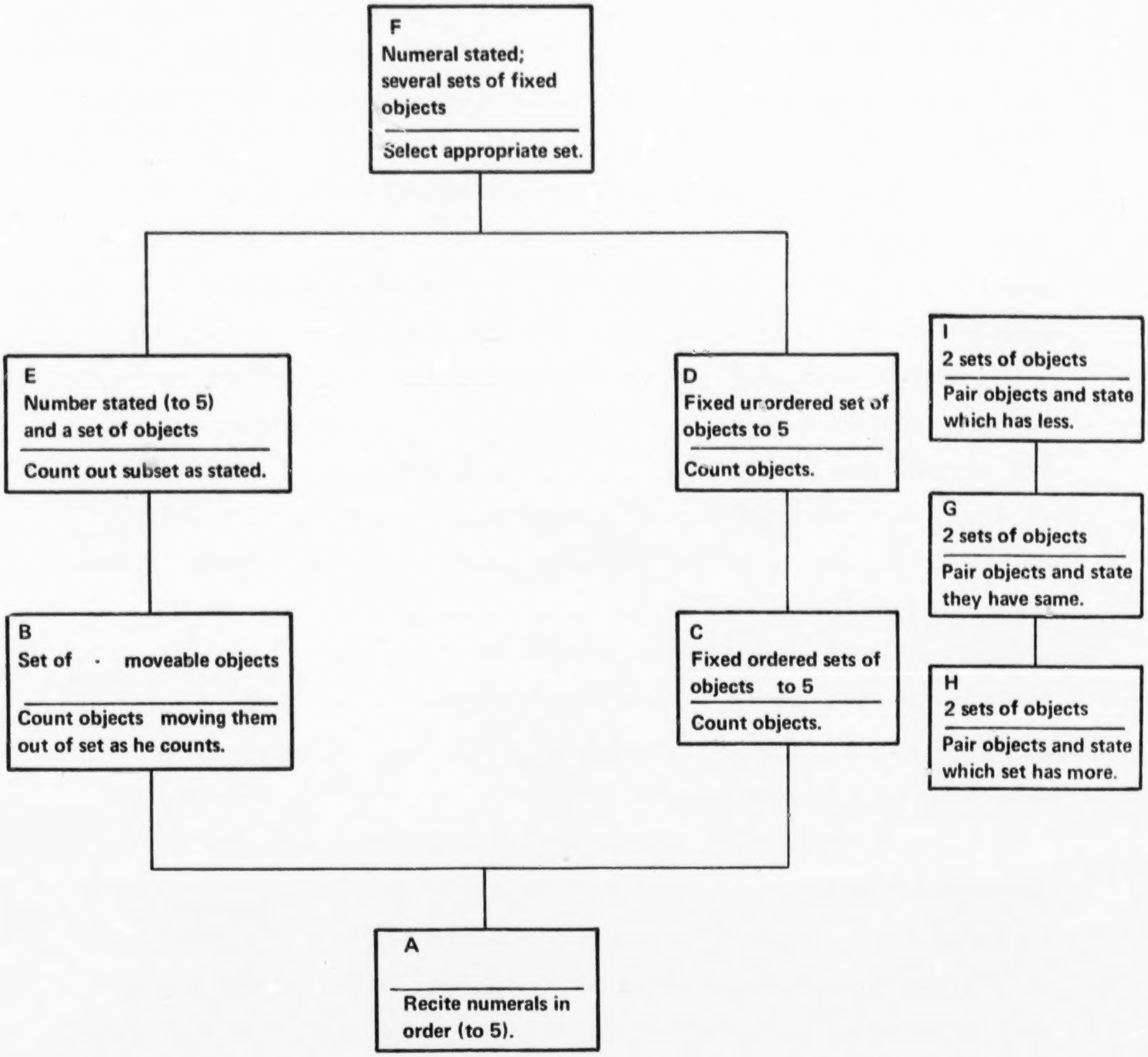

F
Numeral stated; several sets of fixed objects
Select appropriate set.
E
Number stated (to 5) and a set of objects
Count out subset as stated.
D
Fixed unordered set of objects to 5
Count objects.
I
2 sets of objects
Pair objects and state which has less.
G
2 sets of objects
Pair objects and state they have same.
B
Set of · moveable objects
Count objects moving them out of set as he counts.
C
Fixed ordered sets of objects to 5
Count objects.
H
2 sets of objects
Pair objects and state which set has more.
A
Recite numerals in order (to 5).

Figure 2
PEP Introductory Mathematics Curriculum
Quantification Unit 1-Counting (1-5)

integrated two different programs for this crucial area. That the LRDC sequences are so carefully validated doesn't make them inevitable, and respect for the flexibility of human intelligence suggests that while LRDC has found extremely good ways to teach, they should not be viewed as "only" or "best" ways.

5. Must learning hierarchies correspond so slavishly to traditional subject matter areas? If the children need to develop particular generalized communicational, logical, and analytical abilities, it still may be that they can be sponsored by a great variety of content materials. A boy with a passionate interest in science might spend all his days in a science program and emerge with the English, mathematics, history, and so forth that he needs, while a girl with different interests might devote herself to studying history exclusively, yet emerge with similar abilities. The possibility of advancing several complex abilities simultaneously through their interplay in the tasks of a complex, sustained project, as visualized in the rationale of progressive education, may deserve more consideration. Individualized Science shows this kind of thinking, and the "discovery learning" aspect of the PEP program is very promising in this respect.

6. By themselves, can learning hierarchies make explicit the transfer values of what is learned? The PEP team has given definite recognition to transfer values of curriculum objectives when they speak of PEP mathematics providing a basis for the child's continuing experience in mathematics. The fundamental concepts, especially that of number itself, are to be learned in a way that is stable and broad enough to "serve as a conceptual foundation for later work" (Resnick, Wang, & Kaplan, 1970, p. 2). In the adjective "stable" they admit the problem of decay of learning. They want mastery, but mastery of learnings that have been carefully selected for their transfer values. Similarly, the

Perceptual Skills Curriculum has succeeded in showing experimentally that auditory and visual-motor skills can be improved by training, and that these improvements are correlated with success in reading training. This is seen as evidence that specific mechanisms involved in the facilitation of learning are under training. The transfer values of the perceptual-motor program are the bases for its existence. It is noteworthy that this curriculum has made the most open use of traditional trait names in its documentation. An advantage of specifying the traits which learning hierarchies build toward is that traits are widely appreciated as vehicles for the transfer of training to new problems.

7. Can learning hierarchies analysis account for the incidental learnings that occur in any educational program? The pious hope has been expressed that IPI will be adapted to provide the student even greater opportunities for self-direction, "since the ultimate goal of any individualization system is to give the students maximum capability for planning and guiding their own learning" (Lindvall & Cox, 1970, p. 68). This is a statement of a transfer goal, rather than a mastery goal, and it acknowledges that "ultimately" IPI teaches for transfer values of a rather transcendental sort, as do all worthy educational programs. Similarly:

> It is conceivable that individualized instruction will find its major value in attaining not only achievement objectives but other educational goals such as self-direction, self-initiation of one's learning, and the feeling of control over one's learning environment. Success in reaching these outcomes of learning is difficult to measure, but we plan to try to do so (Glaser, 1968, p. 33).

Such outcomes are not included in the carefully specified hierarchies of behavioral objectives, and must therefore be viewed as incidental learnings, although they may be of the greatest importance, as is urged. Criterion-referenced tests will not stretch to cover them. Yet the evaluation requirement to measure incidental learnings has been acknowledged.

"Any program will probably also have some unplanned results, both good and bad. Assessment of the unplanned outcomes may be as useful for both formative and summative evaluation as the measurement of planned outcomes" (Lindvall & Cox, 1970, p. 11). This, from the most extensive monograph on IPI evaluation, would seem to be a concession that invalidates the central argument of the tract, which is that evaluation should be planned in the light of a program's unique goals. Incidental learnings will only be caught if the measurement net thrown out is as broad as possible.

Cronbach has placed the proper emphasis when he dictates: "Consequently, an ideally suitable battery for evaluation purposes will include separate measures of all outcomes the users of the information consider important" (Cronbach, 1971, p. 460). Ironically, he cites Glaser and Klaus (1962, p. 436) for their point that it would be "fundamentally wrong" for an evaluation testing battery to be confined to the content of the instructional program, and he concludes: "The universe pertinent in summative evaluation is the universe of tasks graduates are expected to perform" (p. 460). Years ago, Tyler enunciated what should be basic doctrine in educational psychology.

> Any learning situation has multiple outcomes. While the child is acquiring information, knowledges, and skills, there is also taking place concomitant learnings in attitudes, appreciations, and interests. This view indicates a shift from a narrow conception of subject-matter outcomes to a broader conception of growth and development of individuals (Tyler, 1951, p. 48).

The LRDC staff has not approached the ridiculous posture of that national study of equality of educational opportunity which took a single verbal intelligence test as the sole criterion of educational achievements over the twelve years of public education, but the dedication to learning hierarchies and criterion-referenced tests associated with them may promote the "narrow conception" of which Tyler speaks.

8. Many learning objectives, planned and incidental, that are not abilities measurable by maximum-performance items are important to LRDC. Have any learning hierarchies for non-intellective objectives been created? Given the present state of research, can such be created? The Individualized Science program especially emphasizes such objectives of general personality development, and it is perhaps no accident that it provides few examples of learning hierarchy analyses. It has been suggested that LRDC should turn to the famous Eight Year Study of yesteryear for ideas on how to measure non-intellective outcomes, especially interests (Lindvall & Cox, 1969, p. 187). The recent Nitko and Hsu memo on evaluation (4/20/71) shows awareness of the need to cast a broad net in evaluation research. They list 12 areas of measurement, many of them non-intellective, for which tests are needed. There is a core of concepts and generalizations for the non-intellective facets of personality provided by Project TALENT researches (Lohnes, 1966; Cooley & Lohnes, 1968) to which LRDC could turn for some assistance in planning evaluation from "a broader conception of growth and development of individuals."

These are reasons why learning hierarchies, useful and even essential as they are for curriculum development, are inadequate by themselves for the planning of curriculum evaluation. The same is true of the criterion-referenced tests that accompany them. LRDC needs a supplementary theory of personality structure to organize its summative evaluation program. Traits and factors should be the concepts of that theory. Intellectual power traits should be emphasized, but substantial attention should be given to motives, values, plans, interests, and attitudes as well. Such a descriptive theory of personality will permit LRDC to communicate to the public what it is teaching for, what the transfer values are, and what the incidental outcomes are. A theory of the criteria of educational development should sponsor educational measurement practices that reveal whether pupils are growing "into self-directing adults with free intelligence and responsible orientation toward productive careers" (Lohnes, 1968, p. 119). Only a theory that envisions the man or woman inchoate in the boy or girl can be worthy of the evaluation task.

Dimensions of Educational Treatments

Do different instructional models create different children? Do choices among instructional models represent the differences that make a difference? Cultural anthropology has established that different cultural environments do produce remarkably different children, but in order to show this effect unambiguously, anthropologists have had to compare children from drastically different cultures. It is a nice question whether instructional models can have enough in common to make them competitive in the American educational marketplace and yet produce important differences in children. The nature of the American child and of the American school combine to enforce substantial similarities among competitive curriculum systems. Psychology, sociology, and economics severely constrain the plasticity of the curriculum medium. Curricula, like cold breakfast cereals, will be given unique names by their packagers, but the various contents may not be as unique. People are discrete, unique events, yet personology has not made good psychology. Taxonomy has proven much less effective in psychology than continuous, multivariate measurement along personality dimensions. The solution to the dilemma of findings of unimportant differences in effects of different curricula may be to conceive of curricula as differing in degrees along common dimensions, rather than in kind or type. This would enable substitution of a correlation strategy for the disappointing analysis of variance strategy in evaluation research.

A classic demonstration of the masochism of analysis of variance habits in education is provided by the USOE's Cooperative Reading Studies of 1964-65. These were 27 projects around the country coordinated in general research design, measurement instruments, delineation of treatments, and comparability of data collected. All treatments ran for approximately 140 school days. There were seven distinctly named beginning reading curricula under investigation, although different projects compared various subsets of the methods. A great deal of planning effort went into the definition and specification of the methods, but in the final analysis it appeared that methods given the same name were not implemented the same way from project to project. This led to disagreement among reading experts as to what interpretation could be placed on the findings. The chief coordinators believed, "The use of common measures and ground rules make many comparisons from one study to another possible" (Bond & Dykstra, 1967, p. 2). But elsewhere a reading journal editor had concluded, "No one method should be compared with another because the methods were not sharply and clearly different" (Stauffer, 1966, p. v). Actually, given the large federal investment and the attendant hoopla, comparisons were inevitable, whether or not they were invidious.

The strongest regularity in the data was that organismic inputs measured by the readiness battery, such as general intelligence, auditory perception, and visual discrimination accounted for substantial variance in reading achievement, regardless of the instructional method used. Unfortunately, no treatment-aptitude interactions were discerned. The second strongest source of variance in reading was the projects themselves (which differed in geographic region, urban-rural placement, racial and social class makeup of school populations, et cetera), and even the variance among classes within projects and methods was substantial. There was a distinct Hawthorne effect, in that methods denoted "experimental" within each project yielded better local results than methods locally denoted as "controls." Very little variance was associated with methods. "Even when statistically significant differences were found between treatments, in many cases the differences represented only a month or two of growth according to test norms. There is a question as to whether or not differences of this limited magnitude, although statistically reliable, are really practically significant" (Bond & Dykstra, 1967, p. 7). The architects of the Studies put on the best face they could: "There are indications that reading instruction can be improved . . . . It appears that certain combinations of approaches prove more effective than the use of a certain specified method in isolation" (Bond & Dykstra, 1967, p. 7). The editor is more brutal: "And where does all this leave us? No single approach in these twenty-seven studies has overcome individual differences or eliminated reading disability at the first grade level" (Stauffer, 1966, p. vii).

Individual differences are not meant to be overcome. They are man's blessing, not his curse. A less flagellating and more perceptive comment would be that all seven methods sponsored substantial reading achievements. All seven are good instructional models. That no one appears to be best, nor several appear to be better, in any decisive ways (which will be re-demonstrated by a new analysis later in this essay), either for all children or for some children with special aptitude profiles, is not a criticism of the IM's. It is a failure of research design that leaves the investigators and their audience feeling disappointed. Hopes for positive, new knowledge about beginning reading instruction were doomed when an analysis of variance (with covariance) approach was adopted. Later, the general inadequacy of analysis of variance and covariance for curriculum research will be argued. Here, the point is that reading curricula should not have been conceptualized as discrete, unique "treatments" suitable for levels of a main effect in analysis of variance. All beginning reading curricula must share many common contents, emphases, and techniques, and the ways in which they differ among themselves cannot outweigh these commonalities. Instead of seeking a "best" or some "better" instructional models, research should seek to reveal the correlations between degrees of implementation of various treatment dimensions and degrees of achievements of various types. It should also try to discover whether these treatment-outcome correlations are influenced by non-linear involvements of organismic inputs making moderator effects or treatment-aptitude interactions (TAI's) available. This implies analyses of canonical correlations, multiple partial correlations, and homogeneity of regression systems. But, first, it requires that dimensions of treatment programs be conceptualized, scaled, and measured in school trials.

LRDC has placed great emphasis on the careful description of learning sequences in instructional settings, and has gone so far as to suggest that such descriptions can be the bases of evaluation. The *prima facie* validity of this position is enhanced by the theoretical nature of the descriptions of learning sequences provided by LRDC. The knowledge of learning hierarchies and how to implement them in instruction that LRDC has achieved may be its most important contribution to a science of instruction. This knowledge transcends the usual description of classroom experiences. It is a strong, tested theory of instruction. It is knowledge of process.

Process knowledge can be more important than product knowledge, although they are separate classes of information. People were able to tinker up flying machines, but knowledge of why they flew was necessary for the emergence of the science of aerodynamics and the development of the great flying machines of today. People tinkered up gunpower, but process knowledge of the highest order was a prerequisite to nuclear explosives. Americans worship products. They saw the value of Einstein in the muchroom cloud. Our national underinvestment in pure research is a scandal. LRDC has the problem of showing the educational establishment the long-range value of process knowledge. Specific innovative curricula are the test beds of an evolving instructional model, but the refinement and proving of the instructional model may ultimately represent LRDC's greatest contribution. A science of instruction can be the harbinger of many, varied instructional engineering achievements by many agencies.

Nevertheless, process description is not a substitute for product evaluation. The special usefulness of process thinking in planning a summative evaluation resides in the encouragement it provides to attention to the by-products as well as the intended products of competing curricula. Several curricula may be similarly effective in producing a desired set of competencies, but consideration of their different processes may suggest possibly different influences on attitudes, values, and other incidental learnings.

Description of the instructional model's features cannot be substituted for measurement of the degrees of implementation of dimensions of treatment in the classrooms in which the IM is under trial. Serious slip-ups can and will occur in the implementation of any plans for what many teachers and students in many different places will do. In an evaluation report on the 1967-68 IPI year intended for internal use only, Lindvall (9/11/68) described a study of students required by placement tests to repeat units they had mastered the previous year. Presumably this state of affairs was not intended by the IM, and may be related to poor testing habits when "mastery" was originally claimed. The discouraging thing was that these students often needed as much time to master the unit the second time as they had used "mastering" it the preceding year. This may represent a serious flaw in the theory of instruction, namely gross underestimation of the potential for decay of learning, but it may also and perhaps more likely represent instances of abuse of the curriculum units by these students, and perhaps by their teachers. Another study showed that when the practice of placing students who were doing certain units slowly on "hold" for those units and starting them on advanced units, which is not encouraged by the IM, grew up in one school, many of these students later mastered the obstructed "hold" units rapidly. This is an example of poor IM discipline in a school leading to serendipity. It is nice that some mistakes in the learning hierarchies were discovered, but the present point is that some teachers invented a way of doing with students that was a serious breach of the plan. Another study showed that the prescription-writing behaviors of teachers were based more on their different personalities than on the individual differences among pupils which they were supposed to be reacting to. In short, there are many ways that the intentions of the instructional model may be frustrated in the classroom, and therefore it is essential to have measurement of the actual classroom instructional environment provided.

Although the "Stepping Stones to Reading" program is here considered to be outside the domain of the LRDC IM, the reports on its field testing provide a case study of how extraneous environmental circumstances not specified by the instructional plan may be crucial in determining what students learn. In the city in which the program was most successful, the children had had a strong kindergarten preparation for beginning reading, and the teachers had had previous experience with the Stepping Stones program. Most importantly, however, there was a strong first-grade library reading program, not specified by the experimental program and apparently encouraging reading by contextual clues, something the experimental program wished to discourage, but something which seemed to be a big help (Popp, 1972, in press). Also in this city in which the Stepping Stones program seemed to go very well, "children who were not progressing according to teacher expectations were taken out of the program" (Frankenstein, 1971, p. 27). In a city in which the program did not produce good results a number of factors were disadvantageous, especially a three week teachers' strike (Frankenstein, 1971).

A. W. Astin has been one of the most articulate expressors and ingenious demonstrators of the notion of measuring dimensions of environments. He has recently related this notion explicitly to the problem of evaluation of instruction, in the course of which he has given an excellent general appreciation of the problem:

> Evaluation involves the collection of information concerning the impact of an educational program. While there are many possible uses for such information, it is assumed that the fundamental purpose of evaluation is to produce information which can be used in educational decision making. These decisions may be concerned with the continuation, termination, or modification of an existing program, or with the development and possible adoption of some new program. Whatever the particular decision may involve, evaluation is most likely to produce useful information if it is based on an understanding of the nature of the educational decision-making process itself (Astin & Panos, 1971, p. 733).

The argument presented is that the decision-making process involves some calculus of inputs, operations, and outputs, and therefore, evaluation should relate measures from these three domains of variables to each other. The subjects of evaluative research will have been exposed to different curricula, presumably in a quasi-experimental design allowing in situ research. Under such actual field conditions there must be an imperfect matching of student inputs to the different programs, and it will be better if the treatments are described quantitatively rather than qualitatively, allowing multiple partialling analysis rather than covariance analysis. In short, Astin's paradigm is for relating measured environmental differences among programs to measured output differences in student performances, after partialling for input differences (Astin & Panos, 1971, p. 747). The encouraging thing is that Astin has been very successful in operating this paradigm in his research on effects of college environments, conducted at the American Council on Education.

In a private communication, W. W. Cooley has taken the position that there are three sets of variables to be measured in evaluation studies: (1) community, school, and training variables; (2) the degree of implementation of the model components in each of the classrooms; and (3) student change variables (which almost certainly implies a fourth set of variables, namely organismic input ones). What he does is to elaborate Astin's notion of environmental measures into two categories, his first and second. His first set is similar to the Coleman report independent variables, that is, measures for the whole school or school system in which the trial is conducted. His second set is an inventory of the extent to which each of the engineering features specified by the instructional model is actually present and operating in each classroom. He would use the classroom as the unit of analysis, incidentally (a controversial idea to be discussed later). Examples of implementation variables given are "the establishment of the traveling teacher role, or whether certain subcomponents of the curriculum such as the auditory perceptual skills program are installed in the classroom and made operational" (Cooley, 4/6/71). He claims that reliable and valid measures are presently available for at least a dozen such implementation variables.

One issue of research design raised by Cooley's adding his first set of variables to the Astin paradigm is whether, given good measurement of the degrees of implementation variables in the classrooms, it is really incumbent on the evaluator to measure community and school variables. Cooley believes it is necessary to know what is going on within each sponsoring school system before comparisons are made across sponsors. This is like needing to know the different circumstances for the trials of Stepping Stones in the two cities. It is true that classroom environments will always be correlated with community, home, and general school environments. But isn't it a problem for school sociology to create a theory of those relationships? Things like teachers' strikes and systematic library programs impinge on and help to explain differences among classroom environments. Instructional models have to be robust, but perhaps they can't be asked to be Gargantuan. With Astin and Panos, this paper takes the position that the special task for the evaluation study is to relate carefully measured dimensions of the immediate instructional environment (i.e., in classrooms) to measures of student achievements. Just as classroom environments are correlated with larger environments, so are student organismic inputs, but again the argument is that it is a task for general educational psychology to account for these correlations. The evaluation study should be allowed to make careful measurement of these input differences and appropriate statistical adjustments for them, without being required to explain their distribution. Cooley speaks of what is beyond control, but the position here is that the evaluator should be permitted to assess that which is beyond control without having to explain it. If evaluation is research with a special focus and purpose, it must also be research with special limits.

The crux of the problem is to create a theory of the dimensions of instructional treatments. The recent symposium on evaluation of instruction began auspiciously with talk by one of its editors about the need "to learn how social and other characteristics of the environments of instruction interact with students to mediate changes in their behavior" (Wittrock & Wiley, 1970, p. 16), but failed of its purpose primarily because the participants steadfastly refused to discuss the measurement of environments. The major section of the symposium titled "Instructional

Variables" dealt exclusively with criterion variables! One paper on cost-effectiveness contributed the concept of "manipulatable characteristics" of educational systems, but did not discuss it in useful fashion. What the august symposium skirted is what LRDC should accept as a definite part of its preparation for evaluation studies, namely the creation of a theory and measurement technology for the domain of treatment variables.

What the rubrics and methods, or even the major categories, of such a theory should be is beyond the scope or competence of this essay to propose. Some possible categories that come to mind are:

1. Goals, purposes, and questions provided to students.
2. Materials provided to students.
    a. Structured telling and showing stuff.
    b. Raw, plastic doing stuff.
    c. Incoherent, discovery stuff.
3. Schedules, sequences, calendars, and clock times.
    a. As planned.
    b. As actually happen.
4. Degrees of prescription and control versus autonomy.
5. Rewards and punishments.
6. Teacher behaviors.
7. Testing and assessing devices.
8. Distractions and accidents.
9. Costs and budgets.

Organismic Inputs

Glaser, in private communication: "Treatment effects are easier to get in first grade than in sixth grade. How do you explain this?"

The USOE Cooperative Reading Studies showed they can be very difficult to get in first grade, but the comment is correct. So much of the variance in criterion performances of students after an instructional period of weeks or months will be accounted for by the correlated variance in input tests taken before the period began that the residual criterion scores will show little variance, and their reliability will be low, so that much of the variance they do possess will be error or stochastic variance. This troublesome phenomenon plagues first-grade curriculum studies, but as the grades pass the situation is made worse by the cumulative nature of human learning, which dictates that as children grow older their new learning is increasingly correlated with their previous knowledge. The mind is a filter that controls what it admits to itself, and as it matures its control increases. You may take pains to tell a person what you want to, but you should expect him to take pains to hear what he wants to, and the older he is the more potent this censorship mechanism. Students are not passive receivers. Their minds defend themselves and seek their own aggrandizements, the moreso as they age. In part, this accumulative filtering is automatic and unintelligent, since a person can only learn what he has aptitude or potential to learn, regardless of the demands placed on him. But, it is a common mistake to overestimate the efficacy of the rewards and punishments used by teachers. As a person matures he is increasingly governed by his own interests, values, phobias, and plans and is more and more able to decide for himself to what extent he will comply with the demands of teachers and curricula. These decisions to learn or not to learn are not random but tend to be predictable from his life history. If they are intelligent they are not despicable, even if they frustrate his teachers, and to the extent that they are failures of potential rather than intentional they are not to be despised. Going up the grades, more learning variance is predictable from measured inputs, and more of what is not lost to such control relations is intentional and not engineered by curriculum arrangements.

Treatment effects are never really easy to get if the competitive instructional programs are all based on substantial teaching experience and loving devotion of teachers to pupils. Human intelligence needs social nurturance, moreso when it is young, but the beauty of this evolutionary supertrait is that its development never depends on a particular bag of tricks. I. t. a. (initial teaching alphabet) may be helpful, but it will never be essential. If all children were provided with class A schools and excellent teachers the utilities of particular curriculum programs such as i. t. a., IPI, and PEP would be marginal increments, important perhaps, but relatively small and probably difficult to estimate.

A recent study which demonstrated both the long-range predictive validities of organismic inputs and the elusiveness of treatment effects was based on an opportunity A. P. Newman had to collate sixth-grade achievement data with the first-grade data records of 230 children who were subjects in the Cedar Rapids, Iowa project of the USOE First Grade Reading Studies (Newman, 1971). In September 1964, 51 heterogeneous first-grade classrooms around the city of Cedar Rapids were randomly assigned to seven different beginning reading curricula, named as follows:

Code	Treatment Name	Sample Size
I	Language	32
II	Letter Names	34
III	Literature	30
IV	Skills Development	42
V	Language/Letter Names	35
VI	Language/Literature	23
VII	Language/Skills Development	34
		N = 230

After administration of a lengthy battery of readiness tests, each teacher organized her class into three reading groups, placing those with Metropolitan Readiness Form A total scores below 64 (national 60th percentile) in the "Low" group. Usually this low group contained about one-third of the class. All of Newman's subjects were Low group children in the first grade. At the end of first grade the criterion battery administered consisted of five tests from the Stanford Achievement Tests Primary I. The collated sixth-grade criterion battery consisted of five tests from the Iowa Tests of Basic Skills (ITBS). Table 4 identifies the seventeen readiness tests and the five first-grade and five sixth-grade achievement tests, and provides the total sample means and standard deviations.

Looking first at the predictability of first-grade achievement in language arts (regardless of method of instruction) from sex, age, and the readiness tests, Newman found that 31 per cent of the generalized variance in the five achievement tests was accounted for by the known variance in the readiness battery. (See Cooley & Lohnes, 1971, pp. 168-176 for an explanation of this "redundancy" estimate.) The redundancy captured by the correlation of .68 between the first canonical factors of the batteries, each of which was a g-type construct, was 25 per

Table 4. Newman study. Group means and standard deviations for 27 variables: first-grade readiness (17), first-grade achievement (5), sixth-grade achievement (5) (N = 230)

Test	Name	Mean	S.D.
1	Sex	1.5	.5
2	Chronological Age	77.3	3.4
3	WISC Verbal	96.2	10.8
4	WISC Performance	101.4	12.0
5	Metropolitan Word Meaning	7.4	2.4
6	Metropolitan Listening	9.1	2.2
7	Metropolitan Matching	5.9	3.2
8	Metropolitan Numbers	10.5	3.3
9	Metropolitan Copy	7.6	3.5
10	Metropolitan Alphabet	6.8	3.6
11	Murphy-Durrell Phonemes	16.5	11.9
12	Murphy-Durrell Caps/Lower Case	24.9	10.9
13	Murphy-Durrell Learning Rate	8.1	3.8
14	Thurstone Pattern Copying	7.5	6.2
15	Thurstone Identical Forms	12.6	8.1
16	Bender Gestalt	8.7	3.3
17	Wepman Auditory Discrimination	9.9	6.2
18	SAT Word Reading	15.2	4.6
19	SAT Paragraph Meaning	11.9	4.7
20	SAT Vocabulary	17.6	5.3
21	SAT Spelling	6.5	4.8
22	SAT Word Study Skills	29.3	7.3

Table 4. (continued)

23	ITBS Vocabulary Total	47.2	15.2
24	ITBS Reading Total	48.7	14.0
25	ITBS Language Total	48.7	15.7
26	ITBS Word Study Total	48.6	12.7
27	ITBS Arithmetic	48.4	12.7

cent, so most of the linkage was first-factors correlation. Especially noteworthy were the high first-factor loadings of all five achievement tests. The overall criterion redundancy of .31 was, of course, severely attenuated by the previous restriction of range of general intelligence in the selection of the subjects.

Next, Newman looked at the predictive validity of the readiness battery over a six year span, and found that 34 per cent of the generalized variance in the five sixth-grade achievement tests was redundant to the readiness battery variance. Again, the first canonical factors were g-type constructs correlating .67 with each other, and the redundancy through first-factors linkage was 28 per cent. The total redundancy of sixth to readiness scores was .34. Despite the fact that these were all lowest-third pupils on the Metropolitan Readiness test, the long-range predictive validity of the readiness battery was substantial, and was at the same level as the one-year validity.

Newman's third canonical analysis was between the five first-grade achievement scores and the five sixth-grade ones. She found that 39 per cent of the sixth-grade achievement variance was accounted for by the first-grade achievement variance, and practically all of the redundancy was due to first-factors linkage. Both first factors were strong g constructions, and they were correlated .72, over six years! Thus do organismic inputs foreshadow educational outcomes, even within the "Low" third of children.

What of the seven different treatments? Nothing in achievement was reliably associated with treatments. All the details of Newman's extensive manova and covariance analyses bore out this absence of relationship. No one method was best, nor were some better. What was apparent was that there was excellent teaching in all the classrooms with all the methods. The readiness outlook for these children was poor, yet by sixth-grade-year most of them were achieving at or above grade level. (Actually a few of these subjects were in fifth grade in the follow-up year, but they all took the sixth-grade ITBS form.)

In her summary Newman recapitulates her own inferences about the structure of achievement in elementary school pupils of low school-entering ability:

> Comparisons of the three canonical correlation analyses revealed a strong structural continuity: a striking persistence and intensification of the g-type factor with each successive canonical. This is particularly noteworthy in view of the attenuation of g which probably took place in first grade.
>
> . . . the differentiation between the first and second factors may lend support to Cattell's theory of crystallized versus fluid intelligence. The first canonical factor, which might represent crystallized intelligence or the "good-worker category," appears, in the sixth-grade analysis, to favor the girls. On the other hand, the second factor, favoring a fluid intellect, suggests that boys more than girls absorb and react spontaneously or creatively to the world around them, but do not respond as much with the teacher-pleasing behavior that, for girls, often results in high scores in such subtests as Spelling and Work Study Skills (Newman, 1971, pp. 56-57).

A theory of dimensions of organismic inputs to education will have to emphasize the same traits that are emphasized in a theory of educational criteria. First and foremost will be general intelligence, represented by a g factor of a suitable readiness or aptitude battery. Second will be, of course, existing knowledge of the content area covered by the instructional program. Then there are related knowledges, differential aptitudes, and, perhaps, dimensions of interests. An area not included in the criterion domain is that of achievement motivation. Social class status of family is often used as a surrogate for direct measurement of this area, but such a strategy is full of pitfalls and there is need for improved technology for direct measures.

LRDC's instructional model operates on the principle that where the student is on entry to instruction in relation to the steps of the learning hierarchy is all the input information that is needed. "Regardless of the way a subject matter is structured, there is usually present some hierarchy of subobjectives indicating that certain performances must be present as a basis for learning subsequent tasks. Absence of the specification of prerequisite competence in a sequence of instruction dooms many students to failure" (Glaser, 1968, pp. 6-7). Granting that this is so, the question remains whether necessary prerequisite competence is also sufficient to insure success on the new tasks. Moving up a learning hierarchy presumably calls for integrative skills that are progressively more demanding. Having the required building blocks from the next lowest level does not guarantee that one has the capacity for generalization or transfer that may also be required to establish the broader, more comprehensive traits or trait components of the new level. Do students reach personal plateaus in hierarchies above which they cannot readily progress even if they have the requisite subordinate competencies? Can such plateaus be predicted from measures other than the mastery measures used as placement tests at LRDC? Is general intelligence the most useful of such supplementary measures? It would seem that LRDC would want to do more definitive research on these questions than what it has done.

That intelligence differences are not abolished by the IM is obvious. Individual differences persist under IPI instruction, although the number of mastery objectives achieved replaces the normative test score as their indicator. Glaser (1968) related that the average number of mathematics units completed by 100 students who had been at Oakleaf for three years was 37, with a standard deviation of 12 and a range of 13 to 73. The correlation between where the student entered the program and the number of units completed over three years was .6, while the correlation between the serial number of the entering unit and that of the final unit was .7. He saw the low correlation of .3 between the California Test of Mental Maturity and both number of units completed in three years and serial number of final unit as indicative of the slight usefulness of IQ-type tests. The argument here is that general intelligence deserves to be scaled as the g-type factor from all indicators of performance, and that when it is it will be a powerful explanatory principle.

One of the difficulties in organizing a theory of organismic inputs to instruction for elementary education is that the research on differential aptitudes has been concentrated on adolescent and adult subjects. Probably there are discriminable traits of auditory and visual perception that are important predictors of beginning schooling success that have not been included in the catalog of adult differential aptitudes. The Perceptual Skills program at LRDC represents a productive approach to researching the nature and import of psycho-motor traits in preschool children and developing testing and teaching procedures for the location and correction of deficits (Rosner, 1969). This program posits a variety of auditory, visual, motoric, and integrative functions for processing concrete information from the senses as prerequisites for abstract thinking. "If the common denominator of a group of experiences is not recognized by the child, he is then incapable of the generalization, integration, and categorization necessary to concept formation" (Rosner, Richman, & Scott, 1969, p. 7). An example of the careful reasoning underlying this work is this analysis of the relationship between hearing and learning:

> The auditory-motor component of the LRDC Perceptual Skills Curriculum is based on the rationale that the child's ability to differentiate the phonemic elements of the spoken language develops as the result of feedback loops between his production and hearing of vocal sounds. As the child accumulates experiences, both his hearing and his vocal control gain in the direction of increased capacities for discrete functioning. The

> ability to sort out the perceptual elements of verbal acoustic information seems vital to the subsequent skill of reliably ordering these elements into the symbolic constructs--words--of the culture. As the capacity to sort, order, and synthesize sounds is acquired, refined and performed more efficiently, the task of reliably relating phoneme and grapheme, as required in learning to read and spell, becomes manageable. The goals, then, of the auditory-motor curricular component are to insure that each child acquires the skills needed for competent analysis and synthesis of the phonemes presented in a beginning reading program, and that his repertoire continues to expand as he progresses through that program (Rosner & Simon, 1970, p. 2).

Many children bring these skills as organismic inputs to beginning schooling, "but there may be an undetermined number of children in our schools who are disabled by perceptual-motor deficits" (Rosner, Richman, & Scott, 1969, p. 7). This commitment of PEP to the diagnosis and treatment of aptitudinal deficits should perhaps become a basic principle of the LRDC IM applicable in all curriculum programs.

Rosner and his colleagues are enough impressed with their data on sensory-motor traits to wonder whether it might be useful to classify children into modality preferences, representing some children as preferring to learn graphemes visually before they learned phonemes for them, and others as preferring to learn phonemes in an auditory mode and then to have the graphic elements related to the sounds they know. What they are hypothesizing is one of the treatment-aptitude interactions that have proven so elusive to researchers. "Individual differences--aptitudes--do exist. Consideration should be given to the design of instructional programs that acknowledge individual differences in perceptual aptitudes, identify them, and teach to the student's weaknesses through his strengths" (Rosner & Simon, 1970, p. 21). This always sounds like a good idea.

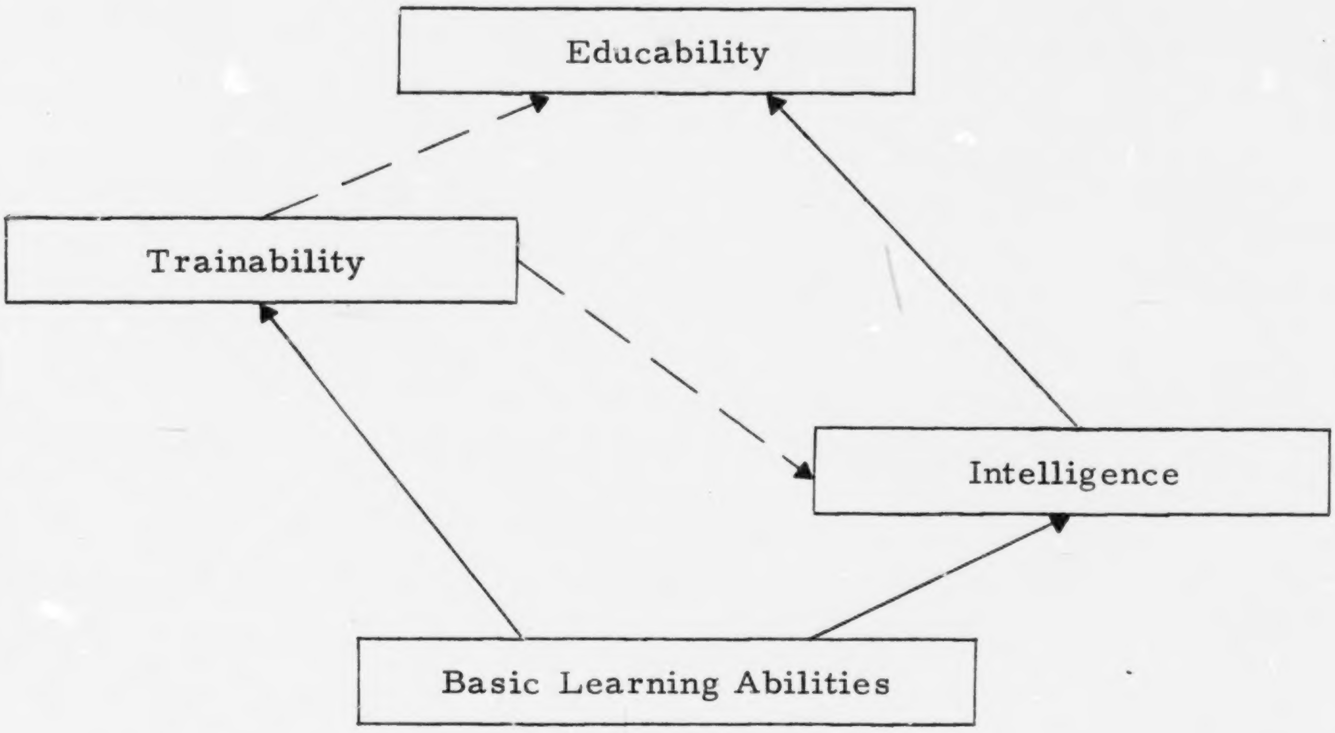
With the nation caught in a miasma of disenchantment with its most talented youth and the federal government caught up in a whirlwind of educational interventions in behalf of the least able, an intriguing question is who benefits most from R&D work such as goes on in the LRDC shop? Do some children who have less benefit from the learning facilitation of general intelligence need the benefits of a strong instructional model more than do children with greater natural ability? Jeanne Chall seems to have been thinking along these lines when she wrote with regard to individual differences:

> a. Children of below-average and average intelligence and children of lower socioeconomic background probably learn better in the end with a code emphasis than with a meaning emphasis, although this advantage does not show immediately.
> b. Children of high mental ability and children of middle and high socioeconomic background appear to gain an immediate advantage from a code emphasis. However, because they are bright they are generally better able to discover sound-letter relationships for themselves. Thus, the differences between results from a meaning or a code emphasis are probably not ultimately as great for them as for average and slow learners and for children of lower socioeconomic background (Chall, 1967, p. 138).

Presumably this eminent professor knows full well that some lower class children are very bright and that some middle and upper class children are not so bright. Since schools have good technology for determining who is bright and who is not without regard to social class status of the child's family, it seems invidious for her to bring in the class issue at all. She must be writing about schools that adopt one beginning reading system for all their children willy-nilly, in which case she points up the need for individually prescribed instruction. At any rate, this is another example of a treatment-aptitude interaction hypothesis. A bit later, she warns that individual differences in general intelligence may override and obscure differential effects of instructional methods, the point that this essay has labored to emphasize:

> Generally, aspects of reading comprehension . . . may not show substantial differences in later years when initial meaning and code programs are compared, since the reader's intelligence and general knowledge put a limit on performance in these areas. However, a code emphasis should still maintain its advantage, even in later years, in those aspects of literacy which depend less on language, intelligence, and experience and more on "reading skill": accuracy in recognizing "unknown" words, accuracy and rate of connected oral reading, rate of silent reading, and some kinds of reading comprehension--e.g., reading for details and following directions. A code emphasis will tend to maintain its early advantage in spelling (Chall, 1967, pp. 138-139).

Arthur Jensen (1968) framed the hypothesis that children of lower intelligence may find special profit in a strong instructional model with the following paradigm:



He said that basic learning abilities "depend relatively little upon mediational processes or specific transfer from previous learning" (Jensen, 1968, p. 33), and gave as examples trial-and-error learning, free recall, serial and paired-associate learning, and digit span. These would appear to be the sort of basic abilities Gagné thought would be found at the base level of learning hierarchies, and they suggest some directions for development of differential aptitude tests for early childhood. He defined educability as "the ability to learn school subjects by means of classroom instruction," and said that to the usual classroom "the learner must bring many developed skills . . . : the voluntary control of attention, the perception of order, self-initiated rehearsal of newly acquired autonomous symbolic mediation, and a host of other processes" (Jensen, 1968, p. 37). These are the sort of mediational processes intelligence provides. What needs to be created for those who lack them is trainability. "Trainability is the ability to acquire knowledge or skill in a situation in which the learner's behavior is under direct, immediate control of the instructor or instructional medium . . . . Focussing of attention, active engagement of the learner, and immediacy of reinforcement are maximized by the instructional technique" (Jensen, 1968, p. 38). Sounds a bit like the LRDC IM. More recently, W. D. Rohwer, Jr. (1971) has addressed himself to some of the same issues made controversial by Jensen's several papers. He differs with Jensen on some important points, and offers his own paradigm which deserves careful study by curriculum designers, but he agrees with Jensen that some of those children who adjust to school so poorly need a special training regimen. He says that "a major objective of curricula in the early years of schooling, especially for low-SES Negro children, should be to assist them in mastering elaborative learning skills, i.e., to actualize children's capacity for imaginative conceptual activity, through concrete, explicit, and specific instructional programs" (Rohwer, 1971, p. 208).

All this is intended to suggest that it would be advantageous for some LRDC programs to take a broader view of the ways children differ among themselves as they enter instruction and of the dynamics of child development. The PEP program certainly has pointed the way. Also, the Individualized Science program has been described in terms of a developmental process that can be analyzed into developmental levels, such that at each level appropriate tasks and behaviors can be defined in relation to the five overall goals of self-direction, co-evaluation, positive affect for science, scientific literacy, and inquiry abilities. The notions of life stages and developmental tasks thereof have played a large role in child psychology. But, the need for the tempering influence of child development theory can be seen in the immodest expectations of the good the LRDC IM will create that occur occasionally in the Center's literature. No instructional model should be expected to provide a general state of grace, from which all goodness flows, so that its students will be superior in such global attributes as sense of worth, creativity, motivation to learn, and total personal adjustment. What is needed is an awareness of the complex etiology of such personality attributes. All theories of child development emphasize, in varying degrees, genetic, familial, early childhood, peer press, and neighborhood determinants of characteristics such as these. School is seen as a late and relatively weak influence. A measurement hazard is created by the fact that schools do drill children in acceptable patterns of language, including self-description. Thus, children may give socially desirable answers to inventory questions administered in school without revealing much of anything about their real personal characteristics.

An LRDC publication that takes a realistic view of child development provides the following analysis of sources of schooling problems.

> Children who are maladjusted in school may be found to be classifiable in at least three subsets:

1. Those with primary emotional disturbances resulting from disturbed interpersonal relationship or adverse psychosocial influences;
2. Those with secondary emotional disturbances stemming from learning disabilities caused by perceptual-motor dysfunction;
3. Those with primary emotional disturbances, accompanied by perceptual-motor deficits (Rosner, Richman, & Scott, 1969, p. 8).

Despite the criticism of the LRDC instructional model for not always acknowledging explicitly the range and depth of organismic inputs to instruction, it should be credited with standing four-square on a foundation of appreciation of individual differences. When the geneticist spoke, LRDC, at least, listened:

> If different individuals, due to their different genotypes, react differently to the same social environmental stimuli, including educational procedures, a method of education which is equally optimal for all individuals can in principle not be developed. . . . The challenge to education appears to me to reside in the problem how to create educational methods and environments which will be optimally adjusted to the needs of unique individuals (Caspari, 1968, pp. 53-54).

Tests for Curriculum Evaluation

For very good reasons, the testing practices within the LRDC curricula, extensive and excellent as they are, provide little basis for external evaluation. Almost paradoxically, an instructional model that strongly emphasizes testing for several purposes in instruction has actually discouraged testing for comparative evaluation. While the IM has been in creation this has been eminently sensible, but with the occurrence of consolidation arrives the need for expansion of the testing program.

Testing in the Instructional Design and Evaluation (IDE) curricula sternly insists on content validity, as expressed by the view that "if IPI test items are not measuring, quite exactly, the specified curricular objectives, they are of no value" (Lindvall & Cox, 1970, p. 24). It is this formative evaluation set that underlies the writing of tests for use within the IM that makes them of limited value for summative evaluation of the IM. The social value of education depends more on transfer of training than it does on the trained repertoire. Because transfer is more important than mastery, tests that measure broad abilities and dispositions are more important than criterion-oriented tests. Construct validity should dominate summative testing as much as content validity dominates formative testing. It is argued that standardized tests are an unfair basis for comparisons of IPI and other IM's: "Standardized tests currently available do not measure student achievement of the type that is defined in the sequences of instructional objectives for IPI" (Lindvall & Cox, 1970, p. 73). But, such tests do not measure specific sequences of behavioral objectives from any program, which is exactly what makes them attractive for comparative studies. They are relevant because they measure the development of general intelligence and, hopefully, levels of development of other broad abilities that are the intervening variables by which education achieves its transfer values.

Some of the notions about using standardized tests in evaluation research that appear in LRDC papers can be criticized. In a memo Klopfer and Champagne (5/6/71) argue that one straightforward approach to summative evaluation of the science program is to select the most widely used science achievement tests and administer them to Individualized Science students. Then comparison with national norms should show that these students do as well as students taught under other IM's. "In addition, we can show, from the data generated in our ongoing formative evaluations, that IDE students are also achieving the specific objectives defined in the several IDE programs." This is both a testing and a data base proposal. Among the problems it ignores are, first, that factor analyses of the most widely used science achievement tests show them to have primarily g loadings with very little special science knowledges loadings; and second, that test publishers' "national" norms are for the most part totally unacceptable as such. American education badly needs the support of a rigorous program of national norming of achievement tests (and other educational measurements). Project TALENT showed some of the ways such a program might be accomplished, but National Assessment, the new look, is a huge step in the wrong direction, and the need persists. Meanwhile, LRDC will have to find better data bases than publishers' norms. Third, how would this scheme estimate the extents to which competitive IM's accomplished the specific objectives of the IDE programs? It is a serious mistake to imagine that other instructional models do not share the grand concerns of the LRDC IM.

A memo from Nitko and Hsu (4/20/71) on evaluation states a strong preference for standardized tests as the measurement devices for evaluation and the intention to create such tests where they are not now available.

They follow with 45 pages of description of some available tests in areas of: (1) mathematics, (2) science, (3) reading, (4) mental growth, (5) reading readiness, (6) creativity, (7) self-reliance, (8) sense of personal worth, (9) social skills, (10) school interest and attitude, (11) motivation to learn, and (12) vocational interests. To try to work with the lexicon and catalog of surface traits scaled by standardized measurements without organizing principles is almost suicidal. TALENT assessed 100 of them, and Guilford has described 222 of them (Guilford, 1960; see also Cooley & Lohnes, 1971, p. 325). Tests will have to be selected for their contributions to the assessment of a factorial model of children's personalities.

A constructive idea put forward by the Nitko and Hsu memo is to create tests in which students are given a new learning task in some subject matter field and access to suitable resource materials. It is hypothesized that IDE students should do a better job with such tasks in terms of: (1) assessing their input competencies, (2) planning their study, (3) using the resources, (4) studying independently, and (5) assessing their outcomes, since these are general capabilities encouraged by the IM. Since test construction is a highly specialized enterprise and the giants in the industry infrequently bring out really new tests, it would be wise for LRDC to be extremely chary of taking on the development of new instruments in support of summative evaluation. Only in areas where there is total conviction that the IM is producing payoffs that cannot be assessed with existing instruments should such undertakings be considered. Even then, every effort should be made to adapt existing curriculum testing procedures to the special demands of summative evaluation.

The later discussion of data bases will examine several cases of LRDC research that has involved standardized tests, but an interesting example of how such tests can contribute to the understanding of a new LRDC diagnostic test is provided by the research on the Auditory Analysis

Test (AAT). This 40-item test "asks the testee to repeat a spoken word, then to repeat it again without certain specified phonemic elements--such as a beginning, ending or medially-positioned consonant sound" (Rosner & Simon, 1970, p. 2). Scores were obtained for 284 children from grades K through 6, and were correlated with Stanford Language Arts and Otis-Lennon IQ, as reported in Table 5. Obviously the AAT has a high g loading, but the authors found encouragement in the partial correlations for the view that the new test also contains unique factor variance that is usefully related to language achievement.

Another example of the vigorous development of special aptitude tests associated with the Perceptual Skills Curriculum is the validation of the Rosner-Richman Perceptual Survey (RRPS), a short form of the Rosner Perceptual Survey made by removing the optometrical and other apparatus items from the latter. The new form for screening with six- to ten-year-olds can be given in fifteen minutes and samples behaviors "within a critical range of sensory-motor processes" (Rosner, Richman, & Scott, 1969, p. 7), including visual-motor function, auditory-motor function, general-motor skills, self-awareness, and integrative function. The authors concluded that their short form is useful for screening purposes, and that it should be operated with cutting scores that allocated about 13 per cent of regular class children, 68 per cent of emotionally disturbed, and 87 per cent of mentally retarded children in a sample to the category of perceptual-motor dysfunction.

The position of this essay is that standardized test batteries can referee the claims of competing instructional models, provided certain problems are met. These problems are dramatized by consideration of the situation of a performance contractor. Should he be allowed to know in advance of his delivery of the educational services the composition of the assessment battery to be used to determine his recompense?

Table 5. Relations among language arts skills, IQ, and Auditory Analysis Test (Rosner & Simon, 1970, p. 13)

Grade	N	Pearson Product Moment Correlations[a]			Partial Correlations[b] between language arts and AAT--IQ held constant
		Language arts[c] and AAT	Language arts[c] and IQ[d]	IQ[d] and AAT	
1	53	.53	.58	.40	.40
2	41	.62	.57	.22	.62
3	37	.84	.76	.67	.69
4	29	.72	.75	.50	.60
5	35	.75	.83	.65	.50
6	39	.59	.86	.64	.10

[a] All correlations significant ($p < .01$, two-tailed) except Grade 2--IQ and AAT (n. s.).

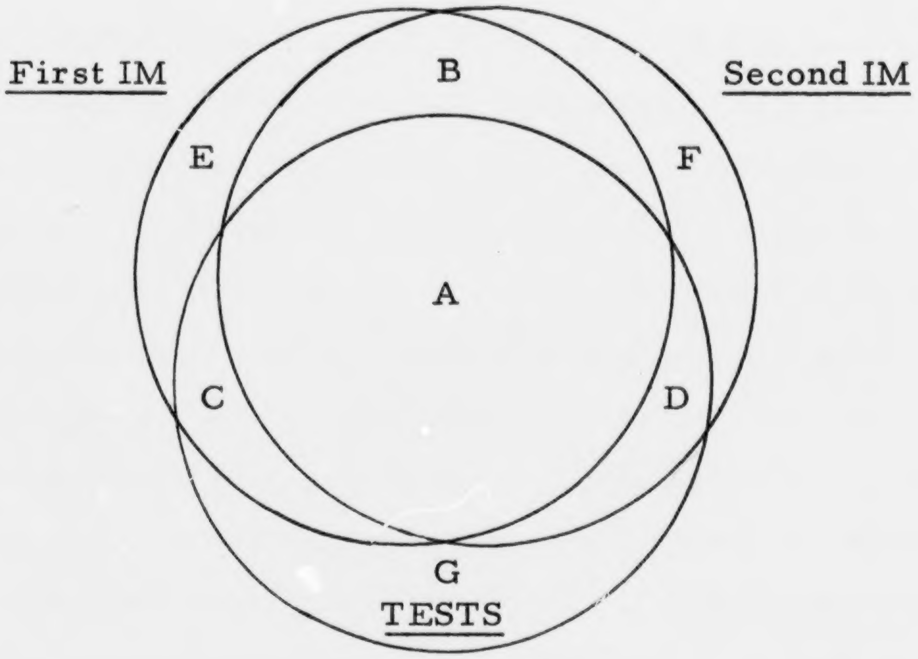
[b] All partial correlations significant ($p < .01$, two-tailed) except Grade 6 (n. s.).

[c] Language arts skills--Stanford Achievement Tests, April 1970--Σ stanines of language arts subtests.

[d] IQ--Otis-Lennon Mental Abilities Test.

Certainly not, if knowing means that he knows what specific standardized tests will be administered. It is too easy to teach for tests in such a situation. But, he deserves to know and indeed must know what factors of mental ability and human motivation the test battery will measure, or, for those who don't like the language of factors, what domains of items will be sampled by the tests and inventories.

The trait development objectives of two IM's will never agree perfectly, nor will any test battery posed by an external evaluator agree in trait sampling with the objectives of either IM perfectly. Cooley (private communication) has suggested the following schematic to represent the complexity of the situation:



The letter areas are intended to indicate specific relationships:

A--Both IM's and the tests agree on traits.

B--Both IM's agree on untested trait objectives.

C--The tests favor special objectives of the first IM.

D--The tests favor special objectives of the second IM.

E--Untested special objectives of the first IM.

F--Untested special objectives of the second IM.

G--The tests sample traits not favored by either IM.

Given this reality, a fair and convincing evaluation can only be possible when the area A dominates the schematic in size. The best way to expedite agreement among the IM's and the tests is to couch the negotiations planning the evaluation in the language of general factors of personality. These factors should be based on orthogonal factor analyses of data matrices, although of course they will come to have surplus meanings. It is amazing how the standard of orthogonality promotes parsimony in factor lexicons. Also, all the factors should be required to possess ubiquitous long-range predictive validities for important real-life placements and adjustments. This will further refine the list.

Research at Project TALENT has yielded a parsimonious factor model of adolescent personality (Lohnes, 1966), with extensive and productive validities for educational and career criteria of early adulthood (Cooley & Lohnes, 1968). The further extension of validity studies for the TALENT measurements is now assured. One possibility for LRDC would be to work out linkages between factors of measurement for elementary education and the TALENT factors. Encouragement and direction for such an effort can be found in Cooley's (1963) overlapping longitudinal study of career development of scientists, which compressed the life-span from grade 5 to "grade 20" (four years beyond baccalaureate) into a five-year research span. His results definitely indicated that factors of ability and interest measured in grade 5 link up with similar factors measured in grade 9, which is where the TALENT measurement began. The logic of overlapping longitudinal research designs was fully explicated by Schaie (1965), who showed how to get estimates of age-to-age relations and of historical time shifts in norms for ages five to twenty from a three-year study with two testings (pre and post) of five samples and one testing (post) of six samples. An adaptation of his sampling scheme to the current LRDC opportunity would be:

First Samples		Second Samples
Pretests (1972)	Posttests (1975)	Posttests (1975)
Age	Age	Age
5	8	5
8	11	8
11	14	11

Very careful preselection of tests and inventories to be given to each cohort would be necessary to hold down the sample size requirement because the multivariate methods to be employed need very robust degrees of freedom if capitalization on chance is to be avoided. Probably samples of 500 to 1000 subjects in each cohort would be wanted. It should be possible, however, to combine this research with an inquiry into the comparative effects of two instructional models, so that an actual demonstration of an evaluation study as well as a theory of measurement for evaluation of elementary educational systems would be produced from the three-year study.

Before any such study begins, the Growth Study resources and experience of Educational Testing Service should be carefully reviewed by the LRDC evaluation staff, particularly as factor-analytic research is now underway on those data.

For LRDC to divert resources from the exciting work of curriculum development to launch a new program of evaluation testing research might seem unfortunate, but certainly the products of the LRDC instructional model deserve to be held up to the best possible mirror. That testing rationale for elementary education which can worthily reflect the comparative utilities of LRDC curricula does not now exist. Therefore, it must be created.

Data Analysis

"The major predictive concern in the measurement of learning outcomes is the relationship between proximate and ultimate educational objectives, and this is more of a learning transfer problem than a correlational one" (Glaser & Nitko, 1971, p. 627). This pronouncement is in agreement with the central thesis of the present essay, namely, that the transfer values of the learnings promoted by an instructional model are more at issue in comparative evaluation research than are the immediate learnings themselves. The thesis continues with the assumption that generalized traits, or factors, of personality are the mediating, intervening variables between proximate and ultimate educational objectives. This is where the pronouncement quoted differs with the present thesis. The former leaves the evaluator with no choice but to conduct long-term and extremely costly longitudinal experiments to establish the direct causal linkages of immediate achievements to vital life adjustments. The latter offers the possibility of referring to existing longitudinal studies of human development for evidence that certain generalized traits have certain vital predictive validities, and then conducting research to establish the comparative utilities of instructional models for the sponsorship of those traits in youth. The contention of this essay is that the only practical way to tackle the problem of long-range transfer values of instruction is through correlational studies in which personality factors are employed as criteria. Is it so bad to fall back on one of the major achievements of almost a century of scientific psychology, the field of differential psychology, for evidence to complete the chain of inference in evaluation research?

The position is that the correlational methods of factor, canonical, and discriminant analyses will best serve the needs of evaluation research.

Before discussion of them, however, a few words about the inadequacies of analysis of covariance are in order, inasmuch as it is the classical procedure of curriculum research. Since the criteria of any reasonable curriculum study are bound to be multiple, forming a vector variable, it may be assumed that the classical approach would involve a multivariate analysis of variance (manova) with covariance adjustment for the regression of the criterion vector variable on a set of organismic input measures, comprising a control vector variable. The independent variable would consist at least of two or several randomly assigned treatment categories, or curricula, but could also include some other design categories, such as "sponsors" (i.e., the schools or school systems in which the experiment was conducted). (The important categorical variable of sex is easily and best accommodated in the control vector variable.)

This research strategy is illustrated in a current text (Cooley & Lohnes, 1971, pp. 292-294) by the example of Balomenos' (1961) experimental study comparing the effectiveness of teaching deduction in two content areas of secondary mathematics. The independent variable was random assignment to geometry or algebra content courses. The criterion vector variable consisted of tests of plane geometry, algebra, general mathematics, deduction, and critical thinking. The control vector variable contained tests of language aptitude, nonlanguage aptitude, geometry, algebra, general mathematics, deduction, and critical thinking. He concluded that the only element of the criterion vector variable on which there was a significant difference after covariance adjustments was plane geometry, for which the effect was almost an entire standard deviation in favor of the geometry content course. This implied that the geometry course kept pace with the algebra course in the teaching of algebra, as well as of general mathematics, deduction, and critical thinking. Covariance analysis seemed to have answered the researcher's question.

The issue is whether he asked enough questions of his data. When he was done he didn't know either the strength or the nature of the regression of the criterion vector on the control vector. He didn't know the correlation matrix for the residual criterion vector or its implications for the structure of learnings. In this worthwhile study, more could have been learned from the data.

Manova analysis of covariance can be a blunt and misdirected tool. Misdirected when it concentrates attention on levels of performance (centroids) at the expense of attention to dispersions. Blunt when it extracts unexamined variance by the regression of the criterion vector variable on the control vector variable. This may often be the major part of the criterion variance, yet covariance as usually practiced doesn't state how much variance is extracted, nor does it provide a theory for the controlled parts of the criterion variance, nor does it state a description of the residual parts of the criterion variance. Covariance can encourage atheoretical empiricism or what might be termed conceptual blindness. Both common sense and serendipity can be discouraged: common sense that the major part of criterion variance in human learning studies will be associated with the organismic input variance, and serendipity in the revelation of the structure of the residual variance. Multiple partialling can be as misleading as covariance in these respects. Canonical and factor analyses are required for informative data handling in the evaluation of instructional models. These methods can tell what variance is associated where, what the dimensionality of association is for each explanatory domain, and what factors or principles of explanation can be employed. It is also significant that when classes are the units of analysis the moments of class distributions can be used as observations in canonical and factor analyses.

In a 1967 memo on planning evaluation of IPI, Lindvall and Cox wrote:

> Of course, the major objective of the program is the individualization of instruction. Defining the objective in this way has important implications for evaluation. It means that the basic step in evaluating IPI in terms of pupil achievement must be to examine pupil variability. Do students actually vary in their levels of achievement and in their rates of progress? Is this variability greater than that which is achieved under other methods of instruction?

It is a short step from this focus on variability, which is technically the diagonal elements of the dispersion matrix for a criterion vector variable, to an interest in the entire dispersion, including the correlation terms. Unfortunately, when Lindvall actually reported an example of IPI evaluation (1970) he reverted to comparing centroids of the four schools using IPI mathematics and reading and their four associated comparison schools. He used school medians on the IPI placement tests and means on the four tests of the Iowa Tests of Basic Skills, and found that the IPI schools showed a median mastery level of about one-half school year above that of comparison schools on IPI tests, but the patterns of ITBS means were so varied for schools-within-treatment as to preclude inferences about treatment effects. In a tantalizing way he suggested that "students from the four IPI schools tested ranked in about the same way in achievement on the two tests" (meaning IPI and ITBS mathematics). How informative some correlation matrices based on these data from three grades in eight schools would have been. The chance to find out how the IPI tests loaded on the g factor that saturates the ITBS tests was missed.

Variance is the essence of humanity. To understand variance is to understand something about the human condition. The only interesting thing about group centroids is their variance across groups. Data analysis for evaluation research should explain the variance in criterion performances by relating it to systematic sources. Since the criterion is a vector variable it is generalized variance, or dispersion, that requires explanation. Technically a dispersion is an excessively complicated and messy collection of variances and covariances that defies human comprehension in the raw. Thus the intrusion of the ubiquitous linear function, or factor, of the measurements. Factors simplify. They make the generalized variance comprehensible.

Canonical correlation explains criterion variance by creating maximally correlated factors of the independent vector variable and the criterion vector variable. A special case of canonical analysis is discriminant analysis, which applies when one of the variables is taxonomic. Discriminant analysis will be useful in evaluation research when it is desirable to relate discrete curricula to criterion variance, but if the argument for scaling and measuring dimensions of instructional treatments is accepted, then the general canonical analysis will apply. In either case, since factors simplify at the expense of incomplete representation of the measurement vectors they are defined on, knowing canonical correlations between paired independent and dependent domain factors is not enough. An estimate of the degree of variance overlap between the two measurement vectors is needed. Such a statistic is provided by the redundancy coefficient invented by Stewart and Love (1968) and exposited in Cooley and Lohnes (1971, pp. 170-173). This descriptive index of how much dependent variable variance is explained by the factors of the independent variable is probably the single most useful statistic for evaluation research.

The data for the USOE Reading Studies Second-grade Phase may be used to demonstrate the power of canonical analysis to explain criterion variance. IBM cards for 3956 pupils (2011 boys and 1945 girls) were obtained from the Coordinating Center of the Studies at the University of Minnesota (thanks to the efforts of Marian M. Gray and the cooperation of

Robert Dykstra). These pupils were from 10 of the original 27 projects of the First Grade Studies. They were in 229 different classrooms to which one of five beginning reading instructional methods had been assigned at the start of the first grade, in September 1964. The 20 projects were sprinkled around the United States, so that they differed on regional, urban-rural, racial, and socioeconomic characteristics, but they were not selected from any sampling frame. The five teaching methods were named:

1. Basal.
2. Initial teaching alphabet (i. t. a.).
3. Language experience.
4. Linguistic.
5. Phonic/linguistic.

These were randomly assigned to classes within projects, insofar as they were implemented in the various projects, but no single project implemented all five methods, and one method (the fifth) was implemented in only one project. All children took an eight-test readiness battery (the organismic inputs) at the beginning of first grade, and all were tested on Stanford Achievement tests at the end of first grade (five tests) and the end of second grade (five tests). Thus each subject was from one of two sexes, one of ten projects, one of five methods, and possessed eight readiness scores. These were the variables that might explain the variance in the ten-test criterion vector variable.

Before a canonical analysis could be computed a correlation matrix for all the variables was needed. Since it was essential that the matrix of intercorrelations of the predictors be of full rank and have an inverse, the predictors had to be coded in such a way as to prevent any complete linear dependencies among them. This was done by using one binary code for sex, nine binary-coded variates for project membership (so that possession of nine zeroes indicated membership in the tenth project), and four binary variates for method of instruction. With the eight input tests this added up to 22 independent variates, and with the 10 achievement tests the order and rank of the correlation matrix was 32. The resulting 32nd order correlation matrix was provided to the canonical program.

The total redundancy of the 10 achievement tests given the 22 predictors was .475. However, of the 10 canonical factors of achievement, only the first two contributed more than .005 to the total redundancy, making it clear that a two-factor theory was quite adequate for this data. Actually, the redundancy contribution of the second factor was only .013, as against .449 for the first, g-type factor, so that little would be lost if a one-factor solution were adopted. Not surprisingly, the first factor of the predictors was essentially a g-type construct from the readiness battery. Sex was coded 0 = boy, 1 = girl, and since sex correlated positively with the second predictor factor, it was apparent that the second achievement factor, which was loaded primarily on spelling, was one on which girls did better than boys. The method loadings on the two predictor factors were negligible. The message is that tested language-arts achievement in first and second grade is moderately predictable, but from general readiness, not from method of beginning reading instruction. That slightly more than half the generalized variance of the achievement vector variable is not accounted for by this complex predictor vector variable which incorporates information about sex, project, method, and readiness poses a real challenge to educational researchers. Tables 6 and 7 report the canonical structures and other details for the first five factors in order to show the basis for the judgment that a rank-two solution is adequate.

Table 6. Canonical factor structure for predictors in USOE Reading Studies data (N = 3956)

Variates	Factors I	II	III	IV	V
1 Sex	.10	.37	.01	-.52	.12
2 Project 1	.20	.49	-.37	.28	-.30
3 Project 2	-.03	-.34	.08	-.10	-.35
4 Project 3	.03	.03	.30	.39	.10
5 Project 4	.06	.16	.14	.09	.13
6 Project 5	.05	-.01	.36	.05	-.38
7 Project 6	-.13	-.09	.06	.21	.22
8 Project 7	-.16	.07	-.34	-.23	.13
9 Project 8	.02	-.26	-.33	-.01	.29
10 Project 9	-.13	.09	-.13	-.32	-.34
11 Basal method	-.01	-.14	.08	-.22	-.12
12 I.t.a. method	.12	.22	-.06	.27	-.08
13 Language experience method	-.11	-.19	-.09	-.07	.24
14 Linguistic method	.00	.06	.14	.17	.28
15 Pintner-Cunningham test	.80	-.23	-.05	-.06	-.16
16 Murphy-Durrell Phonemes	.77	-.20	.32	-.12	.00
17 Murphy-Durrell Letter Names	.77	.20	.08	-.03	.32
18 Murphy-Durrell Learning Rate	.55	.16	.05	-.04	.07
19 Thurstone Pattern Copy	.55	.13	-.09	-.12	-.25
20 Thurstone Identical Forms	.44	.03	-.16	-.03	-.09
21 Metropolitan Word Meaning	.62	-.44	-.20	.26	.16
22 Metropolitan Listening	.54	-.29	-.26	-.02	-.03
Predictor variance extracted	.16	.05	.04	.04	.05
Canonical correlation	.82	.46	.34	.25	.23

Table 7. Canonical factor structure for criteria in USOE Reading Studies data (N = 3956)

Variates	Factors I	 II	 III	 IV	 V
1 Word Reading, grade 1	.86	.18	.39	.11	.22
2 Word Meaning, grade 2	.82	-.01	.09	-.02	.21
3 Paragraph Meaning, grade 1	.86	.39	.09	-.04	-.10
4 Paragraph Meaning, grade 2	.86	.09	-.08	-.15	.21
5 Vocabulary, grade 1	.87	-.27	-.25	.14	-.14
6 Spelling, grade 1	.75	.52	-.07	.15	.09
7 Spelling, grade 2	.72	.30	.20	-.37	.12
8 Word Study Skills, grade 1	.85	.09	.12	-.17	.15
9 Word Study Skills, grade 2	.81	-.05	.29	-.30	-.21
10 Language, grade 2	.77	.05	-.12	-.38	.30
Criterion variance extracted	.67	.06	.04	.05	.03
Canonical correlation	.82	.46	.34	.25	.23
Squared canon correlation	.67	.21	.11	.06	.05
Redundancy to predictor canonical factor	.449	.013	.004	.003	.002

Sometimes it is desirable to build a model for data in discrete steps, extracting the criterion variance that is associated with one set of control or independent variables at each step. A model built in steps cannot be as parsimonious as a canonical model, nor can it avoid the fact that if the sets of independent variables are intercorrelated, as they usually are, the amount of explained variance attributed to a source set will depend in part on the place of that set in the order of the steps. Step modeling only makes sense when there is a strong *a priori* argument for a specific ordering of the set of independent variables. The total proportion of generalized criterion variance accounted for by canonical modeling and any step modeling should be the same. That is, redundancy is invariant. The canonical model itself could be built up in steps by adding in a new set of predictors at each step and observing the increase in redundancy. Another approach would be to partial out sets from the full rank matrix sequentially and compute canonical analyses of the residuals each time. Yet another method will be demonstrated here.

This method of modeling which might be termed analysis of generalized variance depends on a definition of redundancy for an arbitrary factoring procedure. The factoring procedure allows factors to be placed through specific variables, or their remaining residuals, sequentially (Overall, 1962; see also Cooley & Lohnes, 1971, pp. 137-143). In this application, factors are placed through the independent variates, one at a time and in a predetermined order based on an *a priori* argument as to their precedence, and the contribution to total redundancy of the dependent vector variable is computed for each factor and sets of factors corresponding to sets of source variates. If the full matrix of correlations of independent and dependent variables is rank and order m, and there are n variates in the independent vector variable and $m-n$ variates in the dependent vector variable, then the rank of the factor model will be n. Since the number m

represents the complete generalized variance of the correlation matrix, n/m is the proportion of generalized variance due to the independent variable, and $(m-n)/m$ is the proportion of generalized variance due to the dependent variable. The object of the analysis is to estimate how much of the latter is redundant to the former. If V is the cumulative proportion of generalized variance extracted by the n factors, then the total redundancy R_d is given by

$$R_d = (V - (n/m))/((m-n)/m)$$

R_d will have the same value as the total redundancy in a canonical analysis of the n independent variates versus the $n-m$ dependent variates.

If n_j is the number of factors extracted for the jth of k sets of sources of variance, where

$$n = \sum_{j=1}^{k} n_j$$

and V_j is the cumulative proportion of variance extracted by the n_j factors, then

$$R_{d_j} = (V_j - (n_j/m))/((m-n)/m)$$

is the redundant proportion of the dependent domain generalized variance associated with the jth set of the independent variable, when the non-orthogonal sets of sources are ordered for modeling as specified in the plan for factoring.

If $V_j < n_j/m$, making $R_{d_j} < 0$, the meaning is that the variance of the n_j predictors in this set is so confounded with that of the previously extracted predictors that this set makes trivial additional contribution to the explanation of variance in the criterion afforded by this model.

Whether or not some $\underline{R}_{d_j}$ are negative, it will be the case that

$$R_d = \sum_{j=1}^{k} R_{d_j}$$

The great advantage of this modeling procedure is that it allows any mix in any desired order of precedence of continuous measurements and dummy variates of classification or design in the dependent variable. It doesn't matter whether the dummy variates are mutually orthogonal (as in a balanced factorial experimental design) or not (as in most survey research). As opposed to manova procedures, this model works with standard score ($\underline{z}$-score) versions of the dependent variates and the criterion variates. This feature seems to be particularly appropriate for the behavioral sciences in which origins and units of measurement are likely to be thoroughly arbitrary.

The analysis of generalized variance can be demonstrated on the USOE Reading Studies Second-grade Phase data, starting with the same 32nd order correlation matrix for which Tables 6 and 7 report canonical analysis. The order of arbitrary factoring of a rank 22 model was

1. 1 sex factor.
2. 9 project factors.
3. 8 readiness tests factors.
4. 4 methods of instruction factors.

Table 8 reports the communalities of all variates for each set of factors, and for the total model, as well as the $\underline{R}_{d_j}$ and the $\underline{R}_d$, which at .475 agreed exactly with that of the canonical analysis. Very little of this was contributed by sex (.018), but the project domain contributed substantially (.209). This is mostly a case of the "where the brains are" phenomenon, although it may be that the Thurstone Pattern Copy test was maladministered in at least one project. Next, the readiness domain accounted for

Table 8. Analysis of generalized variance for USOE Reading Studies data (N = 3956)

Independent Variates	Source Set Communalities				Total Communality
	Sex	Project	Readin's	Method	
1 Sex	1.00	.00	.00	.00	1.00
2 Project 1	.00	1.00	.00	.00	1.00
3 Project 2	.00	1.00	.00	.00	1.00
4 Project 3	.00	1.00	.00	.00	1.00
5 Project 4	.00	1.00	.00	.00	1.00
6 Project 5	.00	1.00	.00	.00	1.00
7 Project 6	.00	1.00	.00	.00	1.00
8 Project 7	.00	1.00	.00	.00	1.00
9 Project 8	.00	1.00	.00	.00	1.00
10 Project 9	.00	1.00	.00	.00	1.00
11 Basal method	.00	.04	.01	.95	1.00
12 I.t.a. method	.00	.12	.01	.87	1.00
13 Language experience method	.00	.17	.01	.82	1.00
14 Linguistic method	.00	.24	.01	.75	1.00
15 Pintner-Cunningham test	.01	.11	.88	.00	1.00
16 Murphy-Durrell Phonemes	.01	.14	.85	.00	1.00
17 Murphy-Durrell Letter Names	.02	.08	.90	.00	1.00
18 Murphy-Durrell Learning Rate	.00	.12	.88	.00	1.00
19 Thurstone Pattern Copy	.00	.28	.72	.00	1.00
20 Thurstone Identical Forms	.01	.06	.93	.00	1.00
21 Metropolitan Word Meaning	.01	.08	.91	.00	1.00
22 Metropolitan Listening	.00	.05	.95	.00	1.00

Table 8. (continued)

Dependent Variates (Stanford Achievement)	Source Set Communalities Sex	Project	Readin's	Method	Total Communality
1 Word Reading, grade 1	.01	.08	.43	.01	.53
2 Word Meaning, grade 2	.01	.05	.40	.00	.46
3 Paragraph Meaning, grade 1	.02	.08	.42	.01	.53
4 Paragraph Meaning, grade 2	.01	.05	.45	.00	.51
5 Vocabulary, grade 1	.00	.05	.48	.00	.53
6 Spelling, grade 1	.01	.10	.32	.01	.44
7 Spelling, grade 2	.03	.05	.31	.00	.39
8 Word Study Skills, grade 1	.01	.05	.43	.00	.49
9 Word Study Skills, grade 2	.01	.05	.40	.00	.46
10 Language, grade 2	.02	.05	.35	.00	.42
Cumulative variance for entire source set of factors based on 32 variates	.037	.347	.389	.063	.836
Redundancy of dependent variates to independent	.018	.209	.304	-.057	.475

the largest segment of redundancy (.304), and finally the methods of teaching beginning reading failed to pull even their own weight (-.057), indicating trivial additional value for this information after sex, project, and readiness were known.

Manova and covariance focus attention on treatment centroids and their variance. The canonical and generalized variance analyses which have been demonstrated here focus attention on the variance of pupils in criterion performances and estimate the value of treatment information in explaining that variance. Treatment information in these models can just as easily be dimensional as categorical. No problems are created by using classes as units of analysis rather than individual students, if that is desired. These seem to have some claim to superiority as data analytic procedures for evaluation research.

One way that the LRDC IM differs from more conventional instructional models is that it places no meaning on the grouping of pupils in classes. For this reason the use of classes as units of analysis makes less sense in the evaluation of the LRDC IM than it would for most IM's. However, in most IDE demonstration schools the architecture and custom will enforce classroom grouping of pupils willy-nilly, and there may be some communcation value to doing analyses that permit talk about what happens to class distributions under IDE treatment that doesn't happen under alternative treatments. Lohnes (1971) has demonstrated the use of R. A. Fisher's four cumulants, or $\underline{k}$ statistics, as measures of the locations and shapes of class distributions on input and output variables. He showed a canonical regression model for four cumulants of second-grade Stanford Paragraph Meaning distributions of 219 classes from the USOE Reading Studies data on twelve cumulants of three readiness tests. The first canonical R was .89 and the second was .67, with a redundancy of criterion cumulants to readiness cumulants of .39 for the two-factor model. His idea of how such information might be employed in evaluation was as follows:

> Given a class which is to receive an educational treatment that requires evaluation, it will be possible to characterize its syntality in terms of the vector of descriptors for the input tests. It will also be possible to predict from that syntality the canonical factors of descriptors for the criterion test. The ways in which the descriptors for the actual criterion test distribution after treatment deviate from these predictions may indicate the treatment impact on the class (Lohnes, 1971, p. 7).

A major demonstration of Glaser's (1968) ETS paper was the display of 13 full page figures, each of which plotted the units completed in mathematics by one student over three years and fitted a linear least squares line to the progress achieved. For the 13 subjects these lines differed substantially in slope, intercept, and goodness of fit, but presumably they were selected for their differences from among the 100 lines for 100 pupils Glaser had available. Would it have been useful to put all 100 lines on a single graph with the computer plotter? One way to indicate the big picture for a collectivity of students would be to provide statistics for the averages and dispersions of slope, intercept, and standard error of estimate. Standardized tests could then be correlated with these statistics across collectivities such as classes or schools, employing the syntality model based on Fisher's cumulants, as discussed above.

This essay has recommended data analyses for searching out main effects of treatments and criterion correlates of treatment dimensions. The biggest piece of open business in instructional research methodology at present is the delineation of procedures for finding and/or showing treatment-aptitude interactions. Glaser has recently said:

> The criterion against which systems for individualized instruction need to be evaluated is the extent to which they optimize the use of different measures of behavior and

> different alternatives for learning in order to provide different instructional paths (Glaser & Nitko, 1971, p. 666).

It is in this ambition that the LRDC and other instructional models have run ahead of the available research. Despite continued talk about treatment-aptitude interactions at LRDC, Dorothy Zorn's current research on homogeneity of regressions appears to be the only clear-cut example of inquiry after this elusive grail. The question may be asked whether interaction demonstrations are crucial to summative evaluation. Presumably any IM that can validate and capitalize upon treatment-aptitude interactions will be improved in its overall performance as a result. It may be best for summative evaluation to concentrate on revealing and explaining main effects of choices of instructional models. Interaction strategies of instruction may be recognized as treatment dimensions in evaluation research.

Data Bases

With at least four major sources of data at its beck, namely the Oakleaf School and its parent school system, the Frick School, the RBS schools, and the Follow-Through schools, there is no shortage of opportunity for LRDC to conduct evaluative research. Rather, the variety of opportunities creates a problem of deciding what kinds of research are best suited to each data base. It may be helpful to describe two kinds of information which evaluation research should provide and to speculate about appropriate data sources for each. The first kind is information showing that the LRDC instructional model gets better results than do competitive models (if, indeed, this is so). Such information will take the form of reliable, important, and unambiguous contrasts between curricula models. It will be "information that allows one to make the kinds of contrasts necessary in the process of deciding among instructional alternatives" (Cooley, 1/6/71 memo). The second kind is information showing how the LRDC instructional model gets its superior results. This information will take the form of correlations between treatment dimensions and achievement dimensions. It will illuminate the ingredients and recipes for successful instruction. The first kind of information is "black boxes" evaluation. LRDC has a black box and so does some other agency. Contrasts are sought that can say which is the better black box. The second kind of information is more abstract. It helps clients to understand what they are choosing. It looks inside the black boxes to see what makes them go the way they do. Since two competent black boxes are likely to have a lot in common and the superior productivity of both is likely to be more impressive than the marginal contrasts afforded, this process information will be needed by a client wanting to make an intelligent decision.

Evaluators should not expect to get much assistance from inferential statistics. Neither of the types of randomization that make statistical inference possible, which are random sampling from existing populations and random assignment to treatments, is readily available in the planning of evaluations. National probability samples of schools such as Project TALENT achieved for survey purposes are not possible in the testing of new curricula. Regional and state probability samples are just about as unlikely. Within a single school district it might be possible to persuade the chief school officer and the school trustees to assign curricula to schools or to classrooms at random, but the morale damage from the disruption of established habits and the imposition of new demands could be crippling. (One of the treatment dimensions that needs to be scaled is how hard instructional models require teachers to work in a milieu of diminishing effort from workers in general.) Is random assignment consistent with the assumptions of the instructional model? If pupils could be randomly assigned there might not be a problem, but usually it is classes or schools that can be had for random assignment. Also, it should be remembered that randomization achieves its statistical purposes best on large samples, which implies more class or school units randomly assigned than usually can be mustered. The famous small-sample experimental designs depend for their effectiveness on extremely homogeneous pools of material for randomization. Classes and schools are noteworthy for their heterogeneity.

The LRDC IM probably has to be evaluated in "natural" settings, that is, in locations where it has arrived under its own steam with no nudging from a stochastic tugboat. Its competitors should be just as comfortable in their berths. The notion of comparison schools is a good one. Contrasts should be sought between competitive IM's located in quite similar community and school situations. Another notion which seems to have eluded the discussants of LRDC evaluation is that of comparison curricula. The LRDC IM is a grand liner, not to be compared to tramp steamers. Worthy competitors should have the intelligence, the excitement, the élan, and the expense of the LRDC educational system. In their own ways they should be first class. LRDC should seek championship matches and avoid races with derelicts. In studies of IM's in their natural habitats complete unambiguousness of contrasts is impossible because organismic inputs will always be correlated with treatments to some extent. The best that can be done is to match communities and schools as well as circumstances permit and then to openly display the unwanted correlations for which statistical adjustments are made in data analyses. Replication should be the rule. The reliability of important contrasts between worthy competitive instructional models should be demonstrated by replication of them, over and over again if possible. The role of statistical procedures in establishing evaluation contrasts is heuristic more than inferential.

Working in the Frick School, the PEP team has produced a persuasive data base for contrasts. They are showing that youngsters who are admitted to the IDE programs soon outperform their older siblings who are enrolled in the old, established programs in the same school. This demonstration is made possible by the rolling installation of the IDE programs in the Frick School, starting with the preschool year and assuming one new grade each year. The evaluation potential of this scheme is so great that it should be used in installing IDE curricula in many schools. Granted that many schools will turn to LRDC for help when their established programs are in deep trouble and when federal assistance for innovative assaults on almost desperate situations is available, so the competition is likely to be in the derelict class, and the youngsters are likely to include a majority who badly need the help of a strong instructional model and might respond to any such model. LRDC will not be criticized for showing by the Frick plan that it is working in schools where it has been invited to help with serious problems, whereas if LRDC arranges contrasts with other derelict educational systems as "controls" it will invite ridicule. That's the way the world goes. Also, opportunities will occur to roll the IDE programs into some schools that already have strong programs and able students, just because there is an itch to try something new, making it possible to point with candor to the results of the Frick evaluation plan in such schools.

Demonstrating their evaluation plan, the PEP team chose the Wide Range Achievement Test (WRAT) as a criterion variable to supplement the PEP mastery tests. PEP first-graders were five months ahead of grade level in arithmetic whereas non-PEP second-graders were five months behind grade level and non-PEP third-graders were nine months behind grade level. They said, "Thus the non-PEP classes at Frick show evidence of a developing 'cumulative deficit' while PEP classes show promise of having broken the cycle by performing strongly in the first grade" (Wang, Resnick, & Schuetz, 1970, p. 21). The demonstration will be stronger, of course, when second- and third-grade data are in for these PEP classes. This Frick plan provides an excellent setting for collecting longitudinal comparison samples according to the scheme of Schaie (1965), as discussed in the Tests for Curriculum Evaluation section of this essay.

The PEP team reported a strong positive manifold for a correlation matrix involving IQ, WRAT, and PEP mastery tests. When enough data are pooled up over several years a factor analysis demonstration of the importance of general intelligence in their data will be desirable. Already they have concluded, "Since the PEP program is designed to teach those cognitive tasks that are typically measured by intelligence tests, the high correlation between I.Q. and mastery levels in PEP can be interpreted as indicating that the PEP program is doing what it proposes to do" (Wang, Resnick, & Schuetz, 1970, pp. 25-26). They hinted at the possibility of using correlational analysis to show that the IM changes the structure of achievements:

> Overall, the data show that in subjects taught in the PEP curriculum, PEP children scored well, while in areas not explicitly taught they did less well. In the non-PEP classes, by contrast, children performed at about the same level on all three subtests. . . . This finding suggests the power of explicit instruction in the PEP model (Wang, Resnick, & Schuetz, 1970, p. 21).

A series of multiple correlation analyses on PEP data showed strong and similar structural relations between organismic inputs and achievement levels for WRAT and PEP mastery criteria. Family socioeconomic indicators were not especially useful. These studies are summarized in Table 9. (For first-grade WRAT Reading, which space excluded from the table, the multiple R was .59 and the correlation of IQ with the prediction function was .80.) Parallel studies using learning rates as criteria showed much lower multiple R's (averaging .30), with IQ the best predictor by wide margins. When sufficient degrees of freedom are available in the data pool, canonical analyses of these variables will be instructive.

The fullest description of the Frick plan for evaluation studies is in a memo by M. C. Wang (May 1971), which charts the comparisons that will become available over the next few years between: (1) PEP classes and non-PEP classes of the same grade but the previous year; (2) PEP classes and non-PEP classes in upper grades of the same year; (3) PEP classes and PEP classes of the same grade but the previous year; and (4) non-PEP classes and non-PEP classes of the same grade but the previous year. Also, a random sample of 50 children of each age group will be followed more intensively, with a heavy testing schedule imposed on it.

Table 9. Structure coefficients for multiple correlation functions on PEP data

	Criteria and Grade						
Predictors[a]	PEP 1 Quantification	PEP 1 Classification	PEP 1 Reading	PEP K Quantification	PEP K Classification	PEP K Perceptual Skills	WRAT 1 Arithmetic
Quantification Entry	.88	.51	.50	.72	.30	.49	
Quantification Gain							.27
Quantification Terminal							.76
Classification Entry	.64	.86	.63	.61	.59	.30	
Classification Terminal							.50
Perceptual Entry	-.01	-.35	.45	.36	.29	.93	
Intelligence Quotient	.63	.53	.66	.82	.97	.26	.77
Years in PEP	-.16	.14	.25				.19
Father's Education	.14	.03	.39	.34	.48	.12	
Father's Occupation	.04	.06	.06	.03	.12	.13	
Multiple R	.68	.62	.58	.56	.46	.64	.51

[a] Blank spaces indicate predictors not entered into equations for functions.

This plan just has to be considered as the way the LRDC program might be rolled into other schools, one grade per year, in the field testing phase.

Customer wants for product contrasts cannot be ignored, and naturally the LRDC staff and sponsors want the prestige and satisfaction of clearcut demonstrations of the superior quality of their products. But there are human and professional needs that transcend these desires. Cooley (1/6/71 memo) has sounded these in writing that "our loyalty must be to the problem of improving the educational process and not to our first approximation to the solution of that problem (nor to the second, third, etc., for that matter)." The possibility of a final curriculum system died with the medieval social system, if not with the sabretooth tiger. What is more to the point is the need for a powerful and constantly improved theory of instruction. We most need to understand the relationships among pupil capacities and dispositions before and after instruction and dimensions of instructional treatments. As enterprises like LRDC create the theory of instruction it will become possible for less exalted and expensive (or perhaps more entrepreneurial) enterprises to capitalize on the new knowledge to engineer many examples of strong and successful curriculum programs.

In this larger quest the real data base for evaluation is the synthesis of information from many sources that comes to existence in the mind of the educational theorist or practitioner. One educational theoretician has recently written about the validation of educational measurements in terms that apply with equal force to the validation of educational treatments:

> Everything said in this chapter has returned to a concern with understanding. The evidence from a single criterion-oriented validation pins down a fact about a particular local situation. That study, like every other study involving the test, helps to

> amplify the picture of what the test means and how it related to the demands of nontest situations. The study, properly examined, also helps one to understand the nature of educational treatments and their psychological requirements. As they accumulate, therefore, criterion-oriented studies play the same role as do other studies pertinent to construct validation. They generate a theory of individual differences and a theory of tasks and situations. On the strength of such understanding one can make reasonable judgments about the design of new measuring instruments. Since these judgments in turn need to be validated, the process of investigation, and therefore the growth of knowledge, never ends (Cronbach, 1971, p. 503).

Test validation and treatment validation differ in emphases, not in kind. The former emphasizes the refinement of a theory of individual differences, while the latter emphasizes the refinement of a theory of tasks and situations that match to and nurture individual differences. The multivariate correlational studies proposed by this essay may not support the direct, causal inferences that some educators desire, but neither do they make impossible demands for rigorously experimental data bases. They are realistic in their data demands and they hold out the possibility of building complex networks of relational discoveries that will mirror the real world of child development and schooling in understandings that are thoroughly adequate for the engineering of better schools.

Summary of Recommendations

This essay began with a list of seven requirements to be met in planning the evaluation of an instructional model. Six of the requirements have been discussed in the sections of this essay, and the recommendations that have been made with respect to them are collected below. The seventh requirement was expressed as the need for a theory of educational development that takes a long view into the varied futures of children and illuminates the possible value judgments regarding what is good for them, without falling into the trap of a narrow dogmatism that is unsuitable for a free society. The essayist has had his run at this subject in other places, and his personal views have glinted obliquely from the preceding pages. To some he may be guilty of dogmatism in his repetitious harping on general intelligence as the supertrait of paramount importance. Many will not agree with the programmatic view of research on intelligence that is the lodestone for the essayist, and that has been delineated by an English educational psychologist thusly:

> How human beings acquire their intelligence, and the extent to which particular kinds of experience are essential, is a central, if not the central problem in psychology (Butcher, 1968, p. 245).

Fortunately LRDC has at hand the surest guarantee against any narrow dogmatism, in the diversity of insight and opinion possessed by its large and gifted staff. So long as it is the custom at LRDC for people to talk to each other openly and to listen and learn from each other there is little reason to fear a lack of breadth of vision. LRDC has never been guilty of singleness of purpose or of approach. To some extent this essay may have implied a greater coherence in the LRDC instructional model and among the curricular projects than actually exists or is desirable. Like the federal union, LRDC is a union of diversities. As with the federal experience, such a union invites conflict and frustration, but promises the ultimate strength of cooperative endeavor. One of the nicest utterances in the LRDC literature speaks of LRDC striving "towards humanistic goals through behavioral objectives" (Beck, 1970). The spirit in which these recommendations are submitted is one of desire to contribute to LRDC's quest for pluralistic, humanistic goods for youth.

The recommendations are:

1. Find a mnemonic for the "new school" the IDE programs comprise.

2. Produce a source book around a master chart of the IDE curricula, organized into curriculum components within curricula, and curriculum units within components. Show developmental levels and hierarchical relations. Provide internal hierarchies for units in the book, keyed to the master chart.

3. Employ levels of curricula representing about a year's work as the smallest feasible treatments for summative evaluation.

4. Acknowledge general intelligence as the paramount criterion of education in elementary schools.

5. Always operationalize general intelligence as a _g_ factor constructed on a vector of measurements. Never allow a single IQ test score to masquerade as a competent scaling of general intelligence.

6. Conduct factor analyses of IDE curriculum mastery tests along with known marker variables to establish the general trait interpretation of criterion-oriented tests, in order that they may play a role in summative evaluation.

7. Provide names for the general traits constructed by learning hierarchies.

8. Review the progressive education rationale for discovery learning and learning through creative projects.

9. Organize a trait and factor lexicon of the transfer values planned for in LRDC curricula and the incidental learnings that may occur. Respect the ordinary language of education and of differential psychology in this lexicon, but transform some of the established connotations of general trait names through factor analytic research, especially with respect to subject matter achievement traits. Emphasize intellectual power traits, but provide substantially for motives, values, plans, interests, and attitudes as well.

10. Go into the business of conceptualizing, scaling, and measuring dimensions of treatment programs, then research the correlations between degrees of implementation of various treatment dimensions and degrees of achievements.

11. Eschew suggestions that IDE programs place pupils in a state of grace from which they are bound to emerge as paragons.

12. Acknowledge the saliency of developed general intelligence as an organismic input to instruction, and phrase a descriptive theory of organismic inputs in the same lexicon assembled as a theory of the criterion domain.

13. Anticipate that organismic inputs will account for the majority of criterion vector variance in evaluation researches.

14. Attempt to adapt or develop useful measures of achievement motivation in children.

15. Research for the existence of individual learning asymptotes or plateaus that represent the limits of transfer or generalization capacity. If these exist, research for predictors of them.

16. Acknowledge that the overall, long-range distribution impact of the IM should be to exacerbate individual differences, not to reduce them. This principle is a necessary corrective to the emphasis on the mastery notion to prevent a dangerous distortion in the public philosophy.

17. Clarify the issue of whether it is low intelligence (or low "readiness") children who most need the ministrations of a strong instructional model.

18. Expand the lexicon and measurement technology for differential aptitudes and basic learning abilities of young children.

19. Elaborate a theory of child development in terms of life stages, developmental tasks, developmental dynamics, and trait structure of personality, to which the instructional model can be keyed.

20. Adopt a posture on summative testing that insists upon construct validity as sternly as the posture on formative testing insists upon content validity. Select summative tests for their contributions to the assessment of a factorial model of the personality of children.

21. Be extremely chary of undertaking the development of new instruments for summative evaluation. Adopt or adapt existing instruments as much as possible.

22. Require criterion factors to be approximately orthogonal and to have important long-range predictive validities, in order to avoid being swamped under a host of them.

23. Attempt linkages of criterion factors with Project TALENT's validated factors of adolescent personality, through an overlapping longitudinal testing research (three years in duration).

24. Study in detail the ETS Growth Study resources and experiences.

25. Use methods of multivariate analysis in evaluation researches that explain the criterion variance by relating it to systematic sources. Canonical correlation is the best key to this lock, but try the supplementary step modeling procedure of analysis of generalized variance as well. Emphasize redundancy as the most crucial concept and estimate.

26. Try out the model which represents the syntality of a collectivity of pupils by the first four cumulants of organismic input variables,

fits a strong canonical model between cumulants of inputs and of achievements, and characterizes the impact of a treatment on a group by the ways the actual output cumulants deviate from their predicted values.

27. Distinguish "black box" evaluation from evaluation of a theory of instruction and place primary emphasis on the latter.

28. Do not expect totally unambiguous contrasts between curricula.

29. Subordinate inferential statistics to heuristics in data analyses.

30. Supplement the notion of comparison schools with a notion of comparison curricula, and seek worthy competition for IDE programs in field comparisons.

31. Make wide use of the Frick plan for evaluation in one school as a by-product of rolling installation of IDE programs, one grade per year.

32. Continue the tradition of vigorous discussion of evaluation problems in order that LRDC's group intelligence may converge iteratively in the direction of truth.

References

Astin, A. W., & Panos, R. J. The evaluation of educational programs. In R. L. Thorndike (Ed.), *Educational measurement*. (2nd ed.) Washington, D.C.: American Council on Education, 1971. Pp. 733-751.

Balomenos, R. H. An experimental study comparing the effectiveness of teaching deduction in two content areas of secondary mathematics. Unpublished Ed.D. dissertation, Harvard University, 1961.

Beck, I. Towards humanistic goals through behavioral objectives. In J. Maxwell and A. Tovatt (Eds.), *On writing behavioral objectives for English*. Champaign, Ill.: National Council of Teachers of English, 1970. Pp. 97-105.

Bond, G. L., & Dykstra, R. The cooperative research program in first-grade reading. *Reading Research Quarterly*, 1967, *2*, 5-141.

Boozer, R. F. An investigation of selected procedures for the development and evaluation of hierarchical curriculum structures. Unpublished Ph.D. dissertation, University of Pittsburgh, 1970.

Butcher, H. J. *Human intelligence: Its nature and assessment*. London: Methuen, 1968.

Caspari, E. Genetic endowment and environment in the determination of human behavior: Biological viewpoint. *American Educational Research Journal*, 1968, *5*, 43-55.

Chall, J. *Learning to read: The great debate*. New York: McGraw-Hill, 1967.

Cook, W. W. The functions of measurement in the facilitation of learning. In E. F. Lindquist (Ed.), *Educational measurement*. Washington, D.C.: American Council on Education, 1951. Pp. 3-46.

Cooley, W. W. *Career development of scientists: An overlapping longitudinal study*. Cambridge, Mass.: Harvard Graduate School of Education (Cooperative Research Project No. 436), 1963.

Cooley, W. W., & Lohnes, P. R. Predicting development of young adults. Palo Alto, Calif.: American Institutes for Research, 1968.

Cooley, W. W., & Lohnes, P. R. Multivariate data analysis. New York: Wiley, 1971.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D.C.: American Council on Education, 1971. Pp. 443-507.

Frankenstein, R. A beginning reading program "Stepping Stones to Reading": Summary report. Pittsburgh: Learning Research and Development Center, 1971. (Publication 1971/24)

Gagné, R. M., & Paradise, N. E. Abilities and learning sets in knowledge acquisition. Psychological Monographs, 1961, 75, 14.

Glaser, R. Adapting the elementary school curriculum to individual performance. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1968. Pp. 3-36.

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagné (Ed.), Psychological principles in system development. New York: Holt, Rinehart and Winston, 1962. Pp. 419-476.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington, D.C.: American Council on Education, 1971. Pp. 625-670.

Guilford, J. P. Personality. New York: McGraw-Hill, 1960.

Jensen, A. R. Social class, race and genetics: Implications for education. American Educational Research Journal, 1968, 5, 1-42.

Klopfer, L. E. Individualized science in focus. Unpublished manuscript, Learning Research and Development Center, 1971.

Lindvall, C. M. Pupil achievement in IPI Math. Unpublished manuscript, Learning Research and Development Center, 1970.

Lindvall, C. M., & Cox, R. C. The role of evaluation in programs for individualized instruction. In Educational evaluation: New roles, new means 1969. Chicago: The National Society for the Study of Education, 1969. Pp. 156-188. (Also LRDC Reprint 40)

Lindvall, C. M., & Cox, R. C. Evaluation as a tool in curriculum development: The IPI evaluation program. Chicago: Rand McNally, 1970.

Lohnes, P. R. Measuring adolescent personality. Pittsburgh: American Institutes for Research, 1966.

Lohnes, P. R. Reformation through measurement in secondary education. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1968. Pp. 102-121.

Lohnes, P. R. Statistical descriptors of school classes. Unpublished manuscript, State University of New York at Buffalo, 1971.

Lohnes, P. R., & Gray, M. M. Intelligence and the Cooperative Reading Studies. Reading Research Quarterly, 1972, in press.

Murphy, G. Freeing intelligence through teaching. New York: Harper, 1961.

Newman, A. P. Longitudinal study of pupils who were underachieving in reading in first grade. Unpublished Ed. D. dissertation, State University of New York at Buffalo, 1971.

Overall, J. E. Orthogonal factors and uncorrelated factor scores. Psychological Reports, 1962, 10, 651-662.

Popp, H. M. Test project for the LRDC beginning reading program "Stepping Stones to Reading." Pittsburgh: Learning Research and Development Center, 1972, in press.

Resnick, L. B. Design of an early learning curriculum. Pittsburgh: Learning Research and Development Center, 1967. (Working Paper 16)

Resnick, L. B., Wang, M. C., & Kaplan, J. *Behavior analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum.* Pittsburgh: Learning Research and Development Center, 1970. (Monograph 2)

Reynolds, L. J., Light, J., & Mueller, F. Procedures for IPI mathematics: A management model. Unpublished manuscript, Learning Research and Development Center, 1970.

Rohwer, W. D. Learning, race, and school success. *Review of Educational Research*, 1971, *41*, 191-210.

Rosner, J. *The design of an individualized perceptual skills curriculum.* Pittsburgh: Learning Research and Development Center, 1969. (Working Paper 53)

Rosner, J., Richman, V., & Scott, R. H. *The identification of children with perceptual-motor dysfunction.* Pittsburgh: Learning Research and Development Center, 1969. (Working Paper 47)

Rosner, J., & Simon, D. P. *The Auditory Analysis Test: An initial report.* Pittsburgh: Learning Research and Development Center, 1970. (Publication 1971/3)

Schaie, K. A general model for the study of developmental problems. *Psychological Bulletin*, 1965, *64*, 92-107.

Shaycoft, M. F. *The high school years: Growth in cognitive skills.* Pittsburgh: American Institutes for Research, 1967.

Stauffer, R. G. Editor's page: Some tidy generalizations. *The Reading Teacher*, 1966, *20*, 4.

Stewart, D. K., & Love, W. A. A general canonical correlation index. *Psychological Bulletin*, 1968, *70*, 160-163.

Tyler, R. W. The functions of measurement in improving instruction. In E. F. Lindquist (Ed.), *Educational measurement.* Washington, D.C.: American Council on Education, 1951. Pp. 47-67.

Wang, M. C., Resnick, L. B., & Schuetz, P. R. PEP in the Frick Elementary School: Interim evaluation report of the Primary Education Project, 1968-1969. Pittsburgh: Learning Research and Development Center, 1970. (Working Paper 57)

Wittrock, M. C., & Wiley, D. E. (Eds.) The evaluation of instruction: Issues and problems. New York: Holt, Rinehart and Winston, 1970.